



Zeppelin

week-8



default ▾

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

df
```

FINISHED ▶ ⌵ 📖 ⚙

	data1	data2	key1	key2
0	-1.835609	0.600007	a	one
1	-0.843081	0.225136	a	two
2	0.879670	1.200486	b	one
3	0.122031	-1.385600	b	two
4	-0.914893	0.613466	a	one

Took 0 sec. Last updated by anonymous at March 09 2017, 7:04:36 PM. (outdated)

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

FINISHED ▶ ⌵ 📖 ⚙

	data1	data2	key1	key2
0	-1.835609	0.600007	a	one
1	-0.843081	0.225136	a	two
2	0.879670	1.200486	b	one
3	0.122031	-1.385600	b	two
4	-0.914893	0.613466	a	one

Took 0 sec. Last updated by anonymous at March 09 2017, 7:06:20 PM.

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

FINISHED ▶ ⌵ 📖 ⚙

```
key1
a    -1.197861
b     0.500850
Name: data1, dtype: float64
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:06:43 PM.

week-8

```
%pyspark
means = df['data1'].groupby([df['key1'], df['key2']]).mean()
means
```

FINISHED ▶ ⌵ 📖 ⚙

key1 key2
one two
-0.843081

Zeppelin

b one 0.879670
two 0.122031

week-8

Name: data1, dtype: float64

Took 0 sec. Last updated by anonymous at March 09 2017, 7:10:35 PM.

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

FINISHED ▶ 🔍 📖 ⚙️

```
key2      one      two
key1
a      -1.375251 -0.843081
b       0.879670  0.122031
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:12:07 PM.

```
%pyspark
states = np.array(['Ohio','California','California','Ohio','Ohio'])
years = np.array([2005,2005,2006,2005,2006])
df['data1'].groupby([states,years]).mean()
```

FINISHED ▶ 🔍 📖 ⚙️

```
California 2005    -0.843081
           2006     0.879670
Ohio       2005    -0.856789
           2006    -0.914893
```

Name: data1, dtype: float64

Took 0 sec. Last updated by anonymous at March 09 2017, 7:18:09 PM.

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

FINISHED ▶ 🔍 📖 ⚙️

```
      data1      data2
key1
a      -1.197861  0.479536
b       0.500850 -0.092557
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:19:07 PM.

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

FINISHED ▶ 🔍 📖 ⚙️

```
      data1      data2
key1 key2
a      one -1.375251  0.606737
      two -0.843081  0.225136
b      one -0.879670  0.044866
      two  0.122031 -1.385600
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:21:22 PM.

week-8 | Untitled | Untitled | Untitled | Untitled | Untitled | Untitled | Untitled | Untitled | Untitled | Untitled



Zeppelin

FINISHED ▶ ⌵ 📖 ⚙️

week-8

```
%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, 7:22:15 PM.

FINISHED ▶ ⌵ 📖 ⚙️

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

```
a
      data1      data2 key1 key2
0 -1.835609  0.600007    a  one
1 -0.843081  0.225136    a  two
4 -0.914893  0.613466    a  one
b
      data1      data2 key1 key2
2  0.879670  1.200486    b  one
3  0.122031 -1.385600    b  two
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:25:08 PM.

FINISHED ▶ ⌵ 📖 ⚙️

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

```
a one
      data1      data2 key1 key2
0 -1.835609  0.600007    a  one
4 -0.914893  0.613466    a  one
a two
      data1      data2 key1 key2
1 -0.843081  0.225136    a  two
b one
      data1      data2 key1 key2
2  0.879670  1.200486    b  one
b two
      data1      data2 key1 key2
3  0.122031 -1.385600    b  two
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:27:50 PM.

Zeppelin

FINISHED ▶ ⌵ 📖 ⚙

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

week-8

▶ data1 key1 key2 ⬇ 📄



📄 ⚙ 🔒 default ▼

2	0.879670	1.200486	b	one
3	0.122031	-1.385600	b	two

Took 0 sec. Last updated by anonymous at March 09 2017, 7:29:41 PM.

```
%pyspark  
df.dtypes
```

FINISHED ▶ ⌵ 📖 ⚙

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

Took 0 sec. Last updated by anonymous at March 09 2017, 7:30:20 PM.

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

FINISHED ▶ ⌵ 📖 ⚙

```
{dtype('O'):   key1 key2  
0    a  one  
1    a  two  
2    b  one  
3    b  two  
4    a  one, dtype('float64'):  
0 -1.835609  0.600007  
1 -0.843081  0.225136  
2  0.879670  1.200486  
3  0.122031 -1.385600  
4 -0.914893  0.613466}
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:31:43 PM.

ERROR ▶ ⌵ 📖 ⚙

READY ▶ ⌵ 📖 ⚙