capStone_mar9 Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Unt

Zeppelin

capStone_mar9

FINISHED default

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

Took 0 sec. Last updated by anonymous at March 09 2017, 7:01:52 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, 7:02:30 PM.

FINISHED

```
%pyspark
df
```

	data1	data2	key1	key2
0	-1.187821	1.026705	a	one
1	-0.561947	-0.994244	a	two
2	-0.000275	-0.384746	b	one
3	-0.713079	2.054053	b	two
4	1.203388	-0.418335	a	one

Took 0 sec. Last updated by anonymous at March 09 2017, 7:04:19 PM.

FINISHED

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

Took 0 sec. Last updated by anonymous at March 09 2017, 7:08:14 PM.

FINISHED

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

```
key1
a   -0.182127
b   -0.356677
Name: data1, dtype: float64
```

Took 1 sec. Last updated by anonymous at March 09 2017, 7:08:58 PM.

FINISHED

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

Took 0 sec. Last updated by anonymous at March 09 2017, 7:10:22 PM.

FINISHED

```
%pyspark
means
```

```
key2
one      0.007784
two      0.561947
b   one   -0.000275
    two   -0.713079
Name: data1, dtype: float64
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:10:40 PM.

FINISHED

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

```
key2       one       two
key1
a     0.007784 -0.561947
b    -0.000275 -0.713079
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:11:01 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   -0.561947
            2006   -0.000275
Ohio        2005   -0.950450
            2006    1.203388
Name: data1, dtype: float64
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:14:21 PM.

FINISHED

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

```
         data1     data2
key1
a    -0.182127 -0.128625
b    -0.356677  0.834653
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:15:05 PM.

FINISHED

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

```
              data1     data2
key1 key2
a    one   0.007784  0.304185
     two  -0.561947 -0.994244
b    one  -0.000275 -0.384746
     two  -0.713079  2.054053
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:16:32 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, 7:17:11 PM.

FINISHED

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

```
a
      data1     data2 key1 key2
0 -1.187821  1.026705    a  one
1 -0.561947 -0.994244    a  two
4  1.203388 -0.418335    a  one
b
      data1     data2 key1 key2
2 -0.000275 -0.384746    b  one
3 -0.713079  2.054053    b  two
```

Took 1 sec. Last updated by anonymous at March 09 2017, 7:18:08 PM.

FINISHED

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

```
a one
      data1     data2 key1 key2
0 -1.187821  1.026705    a  one
4  1.203388 -0.418335    a  one
a two
      data1     data2 key1 key2
1 -0.561947 -0.994244    a  two
b one
      data1     data2 key1 key2
2 -0.000275 -0.384746    b  one
b two
      data1     data2 key1 key2
3 -0.713079  2.054053    b  two
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:19:17 PM.

FINISHED

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

```
      data1     data2 key1 key2
2 -0.000275 -0.384746    b  one
3 -0.713079  2.054053    b  two
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:31:10 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, 7:31:32 PM.

Zeppelin

capStone_mar9

default

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 -1.187821  1.026705
1 -0.561947 -0.994244
2 -0.000275 -0.384746
3 -0.713079  2.054053
4  1.203388 -0.418335}
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:32:31 PM.

READY