Lab 09 march

```
%pyspark
from pandas import Series, DataFrame
import pandas as pd
```

FINISHED

Took 28 sec. Last updated by anonymous at March 09 2017, 6:58:35 PM.

```
%pyspark
import numpy as np
```

FINISHED

Took 0 sec. Last updated by anonymous at March 09 2017, 6:58:43 PM.

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

FINISHED

Took 0 sec. Last updated by anonymous at March 09 2017, 7:06:18 PM.

```
%pyspark
df
```

FINISHED

	data1	data2	key1	key2
0	-0.260390	1.168887	a	one
1	-0.333507	-1.797469	a	two
2	-0.419739	-0.096406	b	one
3	0.586107	1.162645	b	two
4	-0.942160	-0.167812	a	one

Took 0 sec. Last updated by anonymous at March 09 2017, 7:06:34 PM.

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

FINISHED

Took 0 sec. Last updated by anonymous at March 09 2017, 7:25:38 PM.

```
%pyspark
grouped
```

FINISHED

```
<pandas.core.groupby.SeriesGroupBy object at 0x1102d43d0>
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:25:51 PM.

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

FINISHED

```
key1
a   -0.512019
b    0.083184
Name: data1, dtype: float64
```

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

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

FINISHED

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

```
%pyspark
means
```

FINISHED

```
key1  key2
a     one    -0.601275
      two    -0.333507
b     one    -0.419739
      two     0.586107
Name: data1, dtype: float64
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:26:33 PM.

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

FINISHED

```
key2       one       two
key1
a    -0.601275 -0.333507
b    -0.419739  0.586107
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:26:44 PM.

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

FINISHED

Took 0 sec. Last updated by anonymous at March 09 2017, 7:26:56 PM.

```
%pyspark
years = np.array([2005,2005,2006,2005,2006])
```

FINISHED

Took 0 sec. Last updated by anonymous at March 09 2017, 7:27:05 PM.

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

FINISHED

```
California  2005   -0.333507
            2006   -0.419739
Ohio        2005    0.162858
            2006   -0.942160
Name: data1, dtype: float64
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:27:16 PM.

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

FINISHED

```
         data1     data2
key1
a    -0.512019 -0.265465
b     0.083184  0.533120
```

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

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

FINISHED

```
              data1     data2
key1 key2
a    one  -0.601275  0.500537
     two  -0.333507 -1.797469
b    one  -0.419739 -0.096406
     two   0.586107  1.162645
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:27:38 PM.

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

FINISHED

```
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, 7:27:51 PM.

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

FINISHED

```
a
      data1     data2 key1 key2
0 -0.260390  1.168887    a  one
1 -0.333507 -1.797469    a  two
4 -0.942160 -0.167812    a  one
b
      data1     data2 key1 key2
2 -0.419739 -0.096406    b  one
3  0.586107  1.162645    b  two
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:28:11 PM.

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

FINISHED

```
a one
     data1     data2 key1 key2
0 -0.26039  1.168887    a  one
4 -0.94216 -0.167812    a  one
a two
      data1     data2 key1 key2
1 -0.333507 -1.797469    a  two
b one
      data1     data2 key1 key2
2 -0.419739 -0.096406    b  one
b two
      data1     data2 key1 key2
3  0.586107  1.162645    b  two
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:28:22 PM.

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

FINISHED

Took 0 sec. Last updated by anonymous at March 09 2017, 7:28:35 PM.

```
%pyspark
pieces['b']
```

FINISHED

```
      data1     data2 key1 key2
2 -0.419739 -0.096406    b  one
3  0.586107  1.162645    b  two
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:28:47 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, 7:28:58 PM.

```
%pyspark

grouped = df.groupby(df.dtypes, axis = 1)
```

FINISHED

Took 0 sec. Last updated by anonymous at March 09 2017, 7:29:08 PM.

```
%pyspark
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'):       data1     data2
0 -0.260390  1.168887
1 -0.333507 -1.797469
2 -0.419739 -0.096406
3  0.586107  1.162645
4 -0.942160 -0.167812}
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:29:21 PM.

```
%pyspark
```

READY