DS-670 Lab 6 - Da...

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

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)})
```

FINISHED

```
%pyspark
df
```

FINISHED

	data1	data2	key1	key2
0	-0.870053	2.180266	a	one
1	-1.522811	1.188494	a	two
2	0.108262	-0.168467	b	one
3	2.293439	-0.337486	b	two
4	-1.506748	0.449045	a	one

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

FINISHED

```
%pyspark
grouped
```

FINISHED

```
<pandas.core.groupby.SeriesGroupBy object at 0x10b2cc590>
```

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

FINISHED

```
key1
a   -1.299871
b    1.200851
Name: data1, dtype: float64
```

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

FINISHED

```
%pyspark
means
```

FINISHED

```
key1  key2
a     one    -1.188401
      two    -1.522811
b     one     0.108262
      two     2.293439
Name: data1, dtype: float64
```

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

FINISHED

```
key2       one       two
key1
a    -1.188401 -1.522811
b     0.108262  2.293439
```

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

FINISHED

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

FINISHED

```
California  2005   -1.522811
            2006    0.108262
Ohio        2005    0.711693
            2006   -1.506748
Name: data1, dtype: float64
```

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

FINISHED

```
         data1     data2
key1
a    -1.299871  1.272602
b     1.200851 -0.252976
```

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

FINISHED

```
              data1     data2
key1 key2
a    one  -1.188401  1.314656
     two  -1.522811  1.188494
b    one   0.108262 -0.168467
     two   2.293439 -0.337486
```

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

FINISHED

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

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

FINISHED

```
a
      data1     data2 key1 key2
0 -0.870053  2.180266    a  one
1 -1.522811  1.188494    a  two
4 -1.506748  0.449045    a  one
b
       data1     data2 key1 key2
2  0.108262 -0.168467    b  one
3  2.293439 -0.337486    b  two
```

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

FINISHED

```
a one       data1     data2 key1 key2
0 -0.870053  2.180266    a  one
4 -1.506748  0.449045    a  one
a two       data1     data2 key1 key2
1 -1.522811  1.188494    a  two
b one       data1     data2 key1 key2
2  0.108262 -0.168467    b  one
b two       data1     data2 key1 key2
3  2.293439 -0.337486    b  two
```

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

FINISHED

```
      data1     data2 key1 key2
2  0.108262 -0.168467    b  one
3  2.293439 -0.337486    b  two
```

```
%pyspark
df.dtypes
```

FINISHED

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

```
%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'):       data1     data2
0 -0.870053  2.180266
1 -1.522811  1.188494
2  0.108262 -0.168467
3  2.293439 -0.337486
4 -1.506748  0.449045}
```

FINISHED