DS-670 Lab 7 - 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	0.430353	-0.386396	a	one
1	0.619496	0.501478	a	two
2	-0.137672	0.823586	b	one
3	-0.032792	0.488601	b	two
4	0.485897	-1.011061	a	one

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

FINISHED

```
%pyspark
grouped
```

FINISHED

```
<pandas.core.groupby.SeriesGroupBy object at 0x10b2bff10>
```

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

FINISHED

```
key1
a    0.511915
b   -0.085232
Name: data1, dtype: float64
```

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

FINISHED

```
%pyspark
means
```

FINISHED

```
key1  key2
a     one     0.458125
      two     0.619496
b     one    -0.137672
      two    -0.032792
Name: data1, dtype: float64
```

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

FINISHED

```
key2       one       two
key1
a     0.458125  0.619496
b    -0.137672 -0.032792
```

```
%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    0.619496
            2006   -0.137672
Ohio        2005    0.198781
            2006    0.485897
Name: data1, dtype: float64
```

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

FINISHED

```
         data1     data2
key1
a     0.511915 -0.298660
b    -0.085232  0.656093
```

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

FINISHED

```
              data1     data2
key1 key2
a    one   0.458125 -0.698729
     two   0.619496  0.501478
b    one  -0.137672  0.823586
     two  -0.032792  0.488601
```

```
%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  0.430353 -0.386396    a  one
1  0.619496  0.501478    a  two
4  0.485897 -1.011061    a  one
b
       data1     data2 key1 key2
2 -0.137672  0.823586    b  one
3 -0.032792  0.488601    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  0.430353 -0.386396    a  one
4  0.485897 -1.011061    a  one
a two       data1     data2 key1 key2
1  0.619496  0.501478    a  two
b one       data1     data2 key1 key2
2 -0.137672  0.823586    b  one
b two       data1     data2 key1 key2
3 -0.032792  0.488601    b  two
```

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

FINISHED

```
      data1     data2 key1 key2
2 -0.137672  0.823586    b  one
3 -0.032792  0.488601    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  0.430353 -0.386396
1  0.619496  0.501478
2 -0.137672  0.823586
3 -0.032792  0.488601
4  0.485897 -1.011061}
```

```
%pyspark
```

FINISHED