Class_8_Capstone

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

FINISHED

Took 10 sec. Last updated by anonymous at March 09 2017, 6:58:28 PM.

```
%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)
})
```

FINISHED

Took 0 sec. Last updated by anonymous at March 09 2017, 6:58:44 PM.

```
%pyspark
df
```

FINISHED

```
  Key1 Key2     data1     data2
0    a  one  0.996424 -0.064252
1    a  two  0.524846 -0.330140
2    b  one  0.962509  1.052085
3    b  two -0.047485 -0.176685
4    a  one  0.201132 -0.518024
```

Took 0 sec. Last updated by anonymous at March 09 2017, 6:58:54 PM.

```
%pyspark
grouped = df['data1'].groupby(df['Key1'])
```

FINISHED

Took 0 sec. Last updated by anonymous at March 09 2017, 7:00:53 PM.

```
%pyspark
grouped
```

FINISHED

```
<pandas.core.groupby.SeriesGroupBy object at 0x111579710>
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:01:04 PM.

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

FINISHED

```
Key1
a    0.574134
b    0.457512
Name: data1, dtype: float64
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:01:24 PM.

```
%pyspark
means = df['data1'].groupby([df['Key1'],df['Key2']]).mean()
```

FINISHED

Took 0 sec. Last updated by anonymous at March 09 2017, 7:04:38 PM.

```
%pyspark
means
```

FINISHED

```
Key1  Key2
a     one     0.598778
      two     0.524846
b     one     0.962509
      two    -0.047485
Name: data1, dtype: float64
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:04:44 PM.

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

FINISHED

```
Key2       one       two
Key1
a     0.598778  0.524846
b     0.962509 -0.047485
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:05:25 PM.

```
%pyspark
states = np.array(['ohio','california','california','ohio','ohio'])
```

FINISHED

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

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

FINISHED

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

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

FINISHED

```
california  2005    0.524846
            2006    0.962509
ohio        2005    0.474470
            2006    0.201132
Name: data1, dtype: float64
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:10:56 PM.

```
%pyspark
df.groupby('Key1').mean()
```

FINISHED

```
         data1     data2
Key1
a     0.574134 -0.304139
b     0.457512  0.437700
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:12:28 PM.

```
%pyspark
df.groupby(['Key1','Key2']).mean()
```

FINISHED

```
              data1     data2
Key1 Key2
a    one   0.598778 -0.291138
     two   0.524846 -0.330140
b    one   0.962509  1.052085
     two  -0.047485 -0.176685
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:13:28 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, 7:14:06 PM.

```
%pyspark
for name,group in df.groupby('Key1'):
    print name
    print group
```

FINISHED

```
a
  Key1 Key2     data1     data2
0    a  one  0.996424 -0.064252
1    a  two  0.524846 -0.330140
4    a  one  0.201132 -0.518024
b
  Key1 Key2     data1     data2
2    b  one  0.962509  1.052085
3    b  two -0.047485 -0.176685
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:15:46 PM.

```
%pyspark
for (K1,K2),group in df.groupby(['Key1','Key2']):
    print K1,K2
    print group
```

FINISHED

```
a one
  Key1 Key2     data1     data2
0    a  one  0.996424 -0.064252
4    a  one  0.201132 -0.518024
a two
  Key1 Key2     data1    data2
1    a  two  0.524846 -0.33014
b one
  Key1 Key2     data1     data2
2    b  one  0.962509  1.052085
b two
  Key1 Key2     data1     data2
3    b  two -0.047485 -0.176685
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:17:15 PM.

```
%pyspark
pieces = dict(list(df.groupby('Key1')))
```

FINISHED

Took 0 sec. Last updated by anonymous at March 09 2017, 7:18:36 PM.

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

FINISHED

```
  Key1 Key2     data1     data2
2    b  one  0.962509  1.052085
3    b  two -0.047485 -0.176685
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:18:48 PM.

```
%pyspark
df.dtypes
```

FINISHED

```
Key1      object
Key2      object
data1    float64
data2    float64
dtype: object
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:20:43 PM.

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

FINISHED

Took 0 sec. Last updated by anonymous at March 09 2017, 7:21:41 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  0.996424 -0.064252
1  0.524846 -0.330140
2  0.962509  1.052085
3 -0.047485 -0.176685
4  0.201132 -0.518024}
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:21:59 PM.

```
%pyspark
```

READY