# lab 8

```
%pyspark
import pandas as pd
import numpy as  np
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  1.651658 -1.124743    a  one
1 -0.234707  1.635632    a  two
2  0.091878 -1.221939    b  one
3 -0.859313 -1.293501    b  two
4 -0.300084  1.778034    a  one
```

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

FINISHED

```
%pyspark
grouped
```

FINISHED

```
<pandas.core.groupby.SeriesGroupBy object at 0x10b74f910>
```

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

FINISHED

```
key1
a    0.372289
b   -0.383717
Name: data1, dtype: float64
```

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

FINISHED

```
%pyspark
means
```

FINISHED

```
key1  key2
a     one     0.675787
      two    -0.234707
b     one     0.091878
      two    -0.859313
Name: data1, dtype: float64
```

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

FINISHED

```
key2       one       two
key1
a     0.675787 -0.234707
b     0.091878 -0.859313
```

```
%pyspark

states=np.array(['Ohio','California','California','Ohio','Ohio'])
years=np.array([2005,2005,2006,2005,2006])
df['data1'].groupby([states,years]).mean()
```

FINISHED

```
California  2005   -0.234707
            2006    0.091878
Ohio        2005    0.396173
            2006   -0.300084
Name: data1, dtype: float64
```

```
%pyspark

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

FINISHED

```
         data1     data2
key1
a     0.372289  0.762974
b    -0.383717 -1.257720
```

```
%pyspark

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

FINISHED

```
             data1     data2
key1 key2
a    one   0.675787  0.326645
     two  -0.234707  1.635632
b    one   0.091878 -1.221939
     two  -0.859313 -1.293501
```

```
%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  1.651658 -1.124743    a  one
1 -0.234707  1.635632    a  two
4 -0.300084  1.778034    a  one
b
      data1     data2 key1 key2
2  0.091878 -1.221939    b  one
3 -0.859313 -1.293501    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  1.651658 -1.124743    a  one
4 -0.300084  1.778034    a  one
a two
      data1     data2 key1 key2
1 -0.234707  1.635632    a  two
b one
      data1     data2 key1 key2
2  0.091878 -1.221939    b  one
b two
      data1     data2 key1 key2
3 -0.859313 -1.293501    b  two
```

```
%pyspark

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

FINISHED

```
%pyspark

pieces['b']
```

FINISHED

```
      data1     data2 key1 key2
2  0.091878 -1.221939    b  one
3 -0.859313 -1.293501    b  two
```

```
%pyspark

df.dtypes
```

FINISHED

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

```
%pyspark

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

FINISHED

```
%pyspark

    
```

FINISHED

```
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  1.651658 -1.124743
1 -0.234707  1.635632
2  0.091878 -1.221939
3 -0.859313 -1.293501
4 -0.300084  1.778034}
```

```
%pyspark
```

READY