ZA - March 9, 2017

```
%pyspark
from pandas import Series, DataFrame
import numpy as np, pandas as pd
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 32 sec. Last updated by anonymous at March 09 2017, 7:52:22 PM.

```
%pyspark
df
```

FINISHED

```
      data1     data2 key1 key2
0 -3.207801  0.600166    a  one
1 -1.383139  0.828877    a  two
2 -1.047227 -1.601445    b  one
3 -0.943130 -0.125619    b  two
4 -0.572208 -1.063021    a  one
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:52:37 PM.

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

FINISHED

```
<pandas.core.groupby.SeriesGroupBy object at 0x109a9d0b8>
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:52:39 PM.

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

FINISHED

```
key1
a   -1.721049
b   -0.995178
Name: data1, dtype: float64
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:52:40 PM.

```
%pyspark

grouped.mean()
```

FINISHED

```
key1
a   -1.721049
b   -0.995178
Name: data1, dtype: float64
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:52:42 PM.

FINISHED

```
%pyspark

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

```
key1  key2
a     one    -1.890004
      two    -1.383139
b     one    -1.047227
      two    -0.943130
Name: data1, dtype: float64
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:52:43 PM.

FINISHED

```
%pyspark

means.unstack()
```

```
key2       one       two
key1
a    -1.890004 -1.383139
b    -1.047227 -0.943130
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:52:45 PM.

FINISHED

```
%pyspark

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

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

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

```
California  2005   -1.383139
            2006   -1.047227
Ohio        2005   -2.075465
            2006   -0.572208
Name: data1, dtype: float64
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:52:47 PM.

FINISHED

```
%pyspark

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

```
         data1     data2
key1
a    -1.721049  0.122007
b    -0.995178 -0.863532
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:52:49 PM.

```
%pyspark

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

FINISHED

```
              data1     data2
key1 key2
a    one  -1.890004 -0.231428
     two  -1.383139  0.828877
b    one  -1.047227 -1.601445
     two  -0.943130 -0.125619
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:52:51 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, 7:52:52 PM.

```
%pyspark

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

FINISHED

```
a
      data1     data2 key1 key2
0 -3.207801  0.600166    a  one
1 -1.383139  0.828877    a  two
4 -0.572208 -1.063021    a  one
b
      data1     data2 key1 key2
2 -1.047227 -1.601445    b  one
3 -0.943130 -0.125619    b  two
```

Took 0 sec. Last updated by anonymous at March 09 2017, 8:00:43 PM.

```
%pyspark

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

ERROR

```
    print group
```

```
Traceback (most recent call last):
  File "/var/folders/g3/8csy0jq52kdf7dwf0g391dk40000gn/T/zeppelin_pyspark-6257002556540180687.
py", line 323, in <module>
    code = compile('\n'.join(final_code), '<stdin>', 'exec', ast.PyCF_ONLY_AST, 1)
  File "<stdin>", line 2
    print name
             ^
SyntaxError: Missing parentheses in call to 'print'
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:56:58 PM.

FINISHED

```
%pyspark

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

```
a one
      data1     data2 key1 key2
0 -3.207801  0.600166    a  one
4 -0.572208 -1.063021    a  one
a two
      data1     data2 key1 key2
1 -1.383139  0.828877    a  two
b one
      data1     data2 key1 key2
2 -1.047227 -1.601445    b  one
b two
     data1     data2 key1 key2
3 -0.94313 -0.125619    b  two
```

Took 1 sec. Last updated by anonymous at March 09 2017, 8:01:52 PM.

FINISHED

```
%pyspark

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

pieces ['b']
```

```
      data1     data2 key1 key2
2 -1.047227 -1.601445    b  one
3 -0.943130 -0.125619    b  two
```

Took 0 sec. Last updated by anonymous at March 09 2017, 8:02:42 PM.

FINISHED

```
%pyspark

df.dtypes
```

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

Took 0 sec. Last updated by anonymous at March 09 2017, 8:03:01 PM.

FINISHED

```
%pyspark

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

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 -3.207801  0.600166
1 -1.383139  0.828877
2 -1.047227 -1.601445
3 -0.943130 -0.125619
4 -0.572208 -1.063021}
```

Took 0 sec. Last updated by anonymous at March 09 2017, 8:03:39 PM.

READY

```
%pyspark
```