

GroupBy



```
%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 ▶ ↻ 📖 ⚙️

Took 0 sec. Last updated by anonymous at March 09 2017, 7:03:42 PM.

```
%pyspark
df
```

FINISHED ▶ ⌵ ⌶ ⚙

| | data1 | data2 | key1 | key2 |
|---|-----------|-----------|------|------|
| 0 | 1.345379 | -0.035684 | a | one |
| 1 | -0.410812 | -0.012237 | a | two |
| 2 | -0.341810 | -0.295223 | b | one |
| 3 | 0.434371 | 0.658631 | b | two |
| 4 | 0.179541 | 1.276192 | a | one |

Took 0 sec. Last updated by anonymous at March 09 2017, 7:03:56 PM.

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

FINISHED

Took 0 sec. Last updated by anonymous at March 09 2017, 7:05:29 PM.

```
%pyspark
grouped
```

FINISHED

```
<pandas.core.groupby.SeriesGroupBy object at 0x1087ea4a8>
```

Took 1 sec. Last updated by anonymous at March 09 2017, 7:05:53 PM.

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

FINISHED

```
key1
a    0.371370
b    0.046281
Name: data1, dtype: float64
```

Took 0 sec. Last updated by anonymous at March 09 2017, 7:06:22 PM.

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

FINISHED ▶ 🔍 📖 ⚙️

Group By

default ▼

FINISHED ▶ ↻ 📖 ⚙

FINISHED ▶ ⌵ ⌶ ⚙

↓

↓

FINISHED ▶ 🔍 📖 ⚙️

↓

GroupBy Zeppelin

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

FINISHED ▶ ⌵ 📖 ⚙️

 ⚙️ 🔒 default ▼

```
      data1      data2 key1 key2
2 -0.341810 -0.295223    b  one
3  0.434371  0.658631    b  two
```



Took 0 sec. Last updated by anonymous at March 09 2017, 7:24:06 PM.

```
%pyspark
df.dtypes
```

FINISHED ▶ ⌵ 📖 ⚙️

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

Took 0 sec. Last updated by anonymous at March 09 2017, 7:24:39 PM.

```
%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  1.345379 -0.035684
1 -0.410812 -0.012237
2 -0.341810 -0.295223
3  0.434371  0.658631
4  0.179541  1.276192}
```



Took 0 sec. Last updated by anonymous at March 09 2017, 7:25:33 PM.

ERROR ▶ ⌵ 📖 ⚙️

```
Traceback (most recent call last):
  File "/var/folders/q0/x9k9wvhj1wlc0nhtjb1qt53c0000gn/T/zeppelin_pyspark-3331020174088462452.py", line 253, in <module>
    code = compile('\n'.join(final_code), '<stdin>', 'exec', ast.PyCF_ONLY_AST, 1)
  File "<stdin>", line 1
    dtypes('object'): key1 key2
                      ^
SyntaxError: invalid syntax
```



SyntaxError: invalid syntax

Took 0 sec. Last updated by anonymous at March 09 2017, 7:26:11 PM. (outdated)

READY ▶ ⌵ 📖 ⚙️



Zeppelin

GroupBy





default ▾