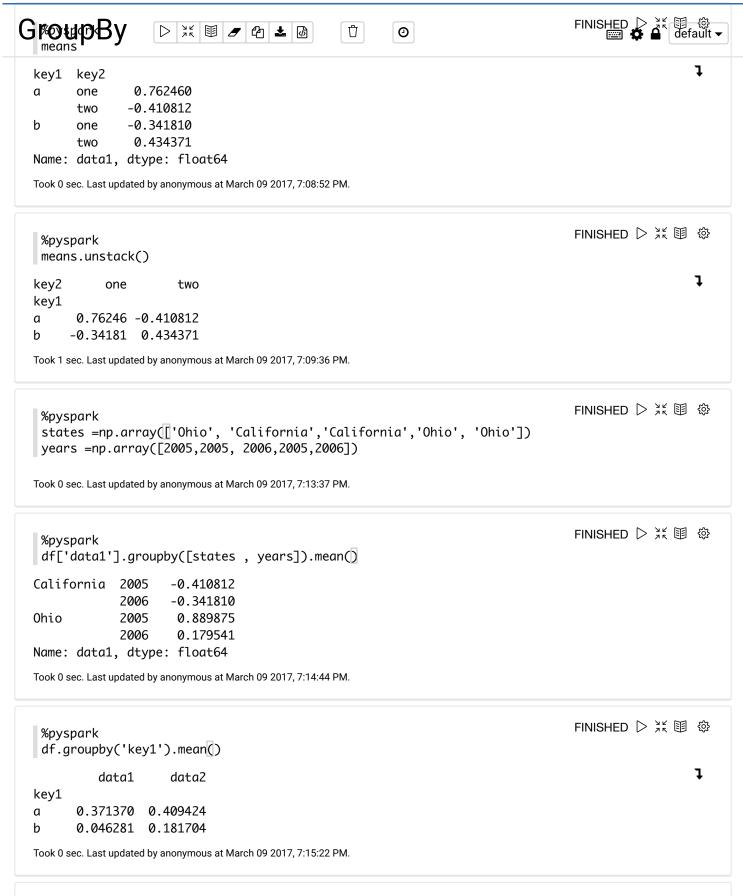
COUPBY Untitled Untit

```
GroupBy
                   ②
                                                                                                 default ▼
                                                                                  FINISHED ▷ ※ 圓 �
  %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)})
 Took 0 sec. Last updated by anonymous at March 09 2017, 7:03:42 PM.
                                                                                  %pyspark
 df
                                                                                                    1
       data1
                  data2 key1 key2
 0 1.345379 -0.035684
                               one
 1 -0.410812 -0.012237
                               two
 2 -0.341810 -0.295223
                               one
   0.434371 0.658631
                            b
                               two
 4 0.179541 1.276192
                            а
                               one
 Took 0 sec. Last updated by anonymous at March 09 2017, 7:03:56 PM.
                                                                                  FINISHED ▷ 💥 🗉 🕸
  %pyspark
  grouped = df['data1'].groupby(df['key1'])
 Took 0 sec. Last updated by anonymous at March 09 2017, 7:05:29 PM.
                                                                                  FINISHED ▷ 光 圓 ۞
  %pyspark
 grouped
                                                                                                    ļ
 <pandas.core.groupby.SeriesGroupBy object at 0x1087ea4a8>
 Took 1 sec. Last updated by anonymous at March 09 2017, 7:05:53 PM.
                                                                                  FINISHED ▷ 💥 🗐 🕸
  %pyspark
  grouped.mean()
                                                                                                    ļ
 key1
      0.371370
      0.046281
 Name: data1, dtype: float64
 Took 0 sec. Last updated by anonymous at March 09 2017, 7:06:22 PM.
                                                                                  FINISHED ▷ 💥 🗐 🕸
  %pyspark
```

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

Untitled Unt



%pyspark df.groupby(['key1','key2']).mean() FINISHED ▷ ♯ 圓 �

```
Untitled Unt
                                            0.762460 0.620254
                            twa 🗖 0,4108<u>12 | 0.012237</u>
                                                                                                                                                                   ②
                                                                                                                                                                                                                                                                                      default ▼
                                       Doj. 341810 -0. 295223
                                             0.434371 0.658631
            Took 0 sec. Last updated by anonymous at March 09 2017, 7:16:17 PM.
                                                                                                                                                                                                                                             FINISHED ▷ 💥 🗐 🕸
              %pyspark
             df.groupby(['key1','key2']).size()
                                                                                                                                                                                                                                                                                               ļ
            key1 key2
                                                      2
                              one
                                                      1
                              two
                              one
                                                      1
            b
                                                      1
                              two
            dtype: int64
            Took 0 sec. Last updated by anonymous at March 09 2017, 7:16:46 PM.
                                                                                                                                                                                                                                             FINISHED ▷ 光 圓 ��
               %pyspark
               for name, group in df.groupby('key1'):
                           print( name)
                           print(group)
                                                                                                                                                                                                                                                                                              ļ
            а
                              data1
                                                            data2 key1 key2
                    1.345379 -0.035684
                                                                                               one
            1 -0.410812 -0.012237
                                                                                              two
                                                                                       а
                     0.179541 1.276192
                                                                                       a one
            b
                              data1
                                                            data2 key1 key2
            2 -0.341810 -0.295223
                                                                                       b
                                                                                               one
                    0.434371 0.658631
                                                                                       b
                                                                                            two
            Took 0 sec. Last updated by anonymous at March 09 2017, 7:20:43 PM.
                                                                                                                                                                                                                                             FINISHED ▷ 💥 🗐 🕸
               %pyspark
               for (k1,k2), group in df.groupby(['key1','key2']):
                           print( name)
                           print(group)
                                                                                                                                                                                                                                                                                               ļ
            b
                              data1
                                                            data2 key1 key2
                    1.345379 -0.035684
            0
                                                                                       а
                                                                                               one
            4
                     0.179541 1.276192
                                                                                               one
                                                                                       а
            b
                              data1
                                                            data2 key1 key2
            1 -0.410812 -0.012237
                                                                                       a two
            b
                                                         data2 key1 key2
                           data1
            2 -0.34181 -0.295223
                                                                                   b one
            b
                                                            data2 key1 key2
                              data1
            3 0.434371
                                                  0.658631
                                                                                       b
                                                                                            two
```

Untitled Unt

```
G FOUDBY: ([ist *addirang(**e/d'))) ①
                                                       ②
                                                                                                default ▼
  pieces ['b']
                                                                                                    ı
       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.
                                                                                 FINISHED ▷ 💥 🗐 🕸
 %pyspark
 df.dtypes
 data1
          float64
 data2
          float64
           object
 key1
 key2
           object
 dtype: object
Took 0 sec. Last updated by anonymous at March 09 2017, 7:24:39 PM.
                                                                                 FINISHED ▷ 💥 🗐 🕸
 %pyspark
  grouped = df.groupby(df.dtypes, axis =1)
 dict(list(grouped))
                                                                                                    ļ
 {dtype('0'):
                 key1 key2
      a one
 1
      a two
 2
      b one
 3
         two
      a one, dtype('float64'):
                                        data1
                                                   data2
   1.345379 -0.035684
 1 -0.410812 -0.012237
 2 -0.341810 -0.295223
    0.434371 0.658631
   0.179541 1.276192}
Took 0 sec. Last updated by anonymous at March 09 2017, 7:25:33 PM.
                                                                                   ERROR ▷ 光 圓 ۞
                                                                                                    1
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
Took 0 sec. Last updated by anonymous at March 09 2017, 7:26:11 PM. (outdated)
```



GroupBy











