IVIGITA Untitled Unti

```
Mar-3
                     ※ 国
                                                  Ů
                                                           ②
                                                                                                                default -
                                                                                               FINISHED ▷ ※ 圓 ��
   %pyspark
   import numpy as np
   import pandas as pd
  Took 0 sec. Last updated by anonymous at March 09 2017, 7:07:35 PM.
                                                                                               FINISHED ▷ 兆 圓 ŵ
   %pyspark
  df = pd.DataFrame({'key1': ['a','a','b','b','a'], 'key2' : ['one','two','one','two','one'], 'data1' : np.random.
  Took 0 sec. Last updated by anonymous at March 09 2017, 7:07:39 PM.
                                                                                               FINISHED ▷ 💥 🗏 🕸
   %pyspark
   df
                                                                                                                    ļ
        data1
                   data2 key1 key2
  0 1.190083 2.016268
                               one
     1.046512 1.580925
                            a two
    0.261075 -0.508882
                            Ь
                               one
  3 -1.517274 1.170868
                            b two
  4 -1.096952 0.393116
                            a one
  Took 0 sec. Last updated by anonymous at March 09 2017, 7:07:55 PM.
                                                                                               FINISHED ▷ ※ 圓 �
   %pyspark
   grouped = df['data1'].groupby(df['key1'])
  Took 0 sec. Last updated by anonymous at March 09 2017, 7:11:29 PM.
                                                                                               FINISHED ▷ ※ 圓 ��
   %pyspark
   grouped
                                                                                                                    1
  <pandas.core.groupby.SeriesGroupBy object at 0x7ffb5c811510>
  Took 0 sec. Last updated by anonymous at March 09 2017, 7:11:48 PM.
                                                                                               FINISHED ▷ ※ 圓 ��
   %pyspark
  grouped.mean()
                                                                                                                    ı
  key1
       0.379881
      -0.628099
  Name: data1, dtype: float64
  Took 0 sec. Last updated by anonymous at March 09 2017, 7:12:16 PM.
                                                                                               FINISHED ▷ ※ 圓 ��
   %pyspark
   means = df['data1'].groupby([df['key1'],df['key2']]).mean()
```

Untitled Unt

```
Mar-3
                                                                                                       FINISHED ♣ ♣ default ▼
                          光 圓
                                                        Ů
                                                                 ②
                                                                                                                             1
     key1
            key2
            one
                     0.046566
                     1.046512
            two
                     0.261075
     b
            one
            two
                    -1.517274
     Name: data1, dtype: float64
     Took 0 sec. Last updated by anonymous at March 09 2017, 7:15:04 PM.
                                                                                                       FINISHED ▷ 光 圓 �
       %pyspark
      means.unstack()
                                                                                                                             ļ
     key2
                 one
                            two
     key1
            0.046566 1.046512
     a
            0.261075 -1.517274
     Took 0 sec. Last updated by anonymous at March 09 2017, 7:15:47 PM.
                                                                                                       FINISHED D 端 圓 繳
       %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
                           1.046512
                   2006
                           0.261075
     0hio
                   2005
                          -0.163595
                   2006
                          -1.096952
     Name: data1, dtype: float64
     Took 0 sec. Last updated by anonymous at March 09 2017, 7:19:56 PM.
                                                                                                       FINISHED ▷ ※ 圓 ��
       %pyspark
       df.groupby('key1').mean()
                                                                                                                             ļ
               data1
                          data2
     key1
            0.379881 1.330103
           -0.628099 0.330993
     Took 1 sec. Last updated by anonymous at March 09 2017, 7:20:48 PM.
                                                                                                       FINISHED ▷ 💥 🗏 🕸
       %pyspark
       df.groupby(['key1','key2']).mean()
                                                                                                                             ļ
                     data1
                               data2
     key1 key2
                 0.046566 1.204692
           one
                 1.046512 1.580925
                 0.261075 -0.508882
                -1.517274 1.170868
           two
Took 0 spc. Last updated by anonymous at March 09 2017, 7:21:36 PM.
```

Untitled Unt

```
FINISHED ▷ 牂 邼 ��
 %pvspark
 df.groupby(['key1','key2']).size()
                  * I
                                              Ů
                              2
                                                      ②
                                                                                                          default ▼
             2
             1
      two
Ь
      one
             1
      two
dtype: int64
Took 0 sec. Last updated by anonymous at March 09 2017, 7:22:14 PM.
                                                                                         FINISHED ▷ 兆 圓 ۞
 %pyspark
 for name, group in df.groupby('key1'):
     print name
     print group
                                                                                                              ļ
а
      data1
               data2 key1 key2
  1.190083 2.016268
                        a one
  1.046512 1.580925
                         a two
 -1.096952 0.393116
                         a one
      data1
               data2 key1 key2
2 0.261075 -0.508882
                         b one
3 -1.517274 1.170868
                        b two
Took 0 sec. Last updated by anonymous at March 09 2017, 7:23:23 PM.
                                                                                         %pyspark
 for (k1,k2), group in df.groupby([['key1','key2']):
     print k1, k2
     print group
                                                                                                              ļ
a one
      data1
                data2 key1 key2
 1.190083 2.016268
                        a one
4 -1.096952 0.393116
                         a one
a two
               data2 key1 key2
      data1
1 1.046512 1.580925
b one
      data1
               data2 key1 key2
2 0.261075 -0.508882
                        b one
b two
                data2 key1 key2
      data1
3 -1.517274 1.170868
                        b two
Took 0 sec. Last updated by anonymous at March 09 2017, 7:24:29 PM.
                                                                                         FINISHED ▷ 光 圓 ⑫
 %pyspark
 pieces = dict(list(df.groupby('key1')))
pieces['b']
                                                                                                              ı
      data1
                data2 key1 key2
2 0.261075 -0.508882
                         b one
3 -1.517274 1.170868
                        b two
```



