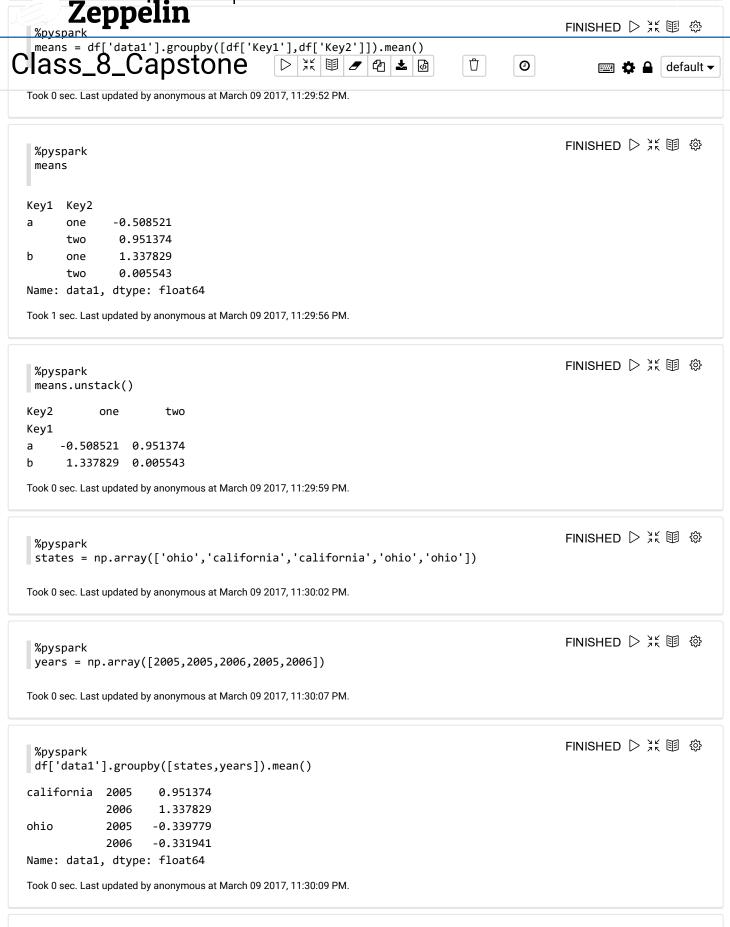
## Class\_o\_caps to reputited untitled unti

```
Class_8_Capstone
                                            Û
                                                                                  ②
                                                                                                         default -
                                                                                         FINISHED ▷ ※ 圓 ��
      %pyspark
      import pandas as pd
      import numpy as np
     Took 45 sec. Last updated by anonymous at March 09 2017, 11:29:34 PM.
                                                                                         FINISHED ▷ ※ 圓 ��
      %pyspark
      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 28 sec. Last updated by anonymous at March 09 2017, 11:29:34 PM.
                                                                                         %pyspark
     df
                     data1
                                data2
       Key1 Key2
          a one -0.685101 -1.112390
                  0.951374 1.871486
             two
                 1.337829 1.187213
     3
             two 0.005543 -0.820762
             one -0.331941 0.084290
     Took 0 sec. Last updated by anonymous at March 09 2017, 11:29:39 PM.
                                                                                         FINISHED ▷ ※ 圓 ��
      %pyspark
      grouped = df['data1'].groupby(df['Key1'])
     Took 0 sec. Last updated by anonymous at March 09 2017, 11:29:43 PM.
                                                                                         FINISHED ▷ ※ 圓 ��
      %pyspark
     grouped
     <pandas.core.groupby.SeriesGroupBy object at 0x00000000000600F400>
     Took 0 sec. Last updated by anonymous at March 09 2017, 11:29:45 PM.
                                                                                         FINISHED ▷ ♯ 圓 墩
      %pyspark
     grouped.mean()
     Key1
         -0.021889
          0.671686
     Name: data1, dtype: float64
Took 1 sec. Left up deted by anonymous at March 09 2017, 11:29:49 PM.
```

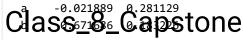
UIASS\_O\_UAPSIUITE Untitled Unt



%pyspark df.groupay('key1').mean() FINISHED ▷ 圓 ۞

## UIASS\_O\_UAPSIUIT Untitled Unti

Key1













Took û sec. Last updated by anonymous at March û9 2017, 11:30:14 PM.

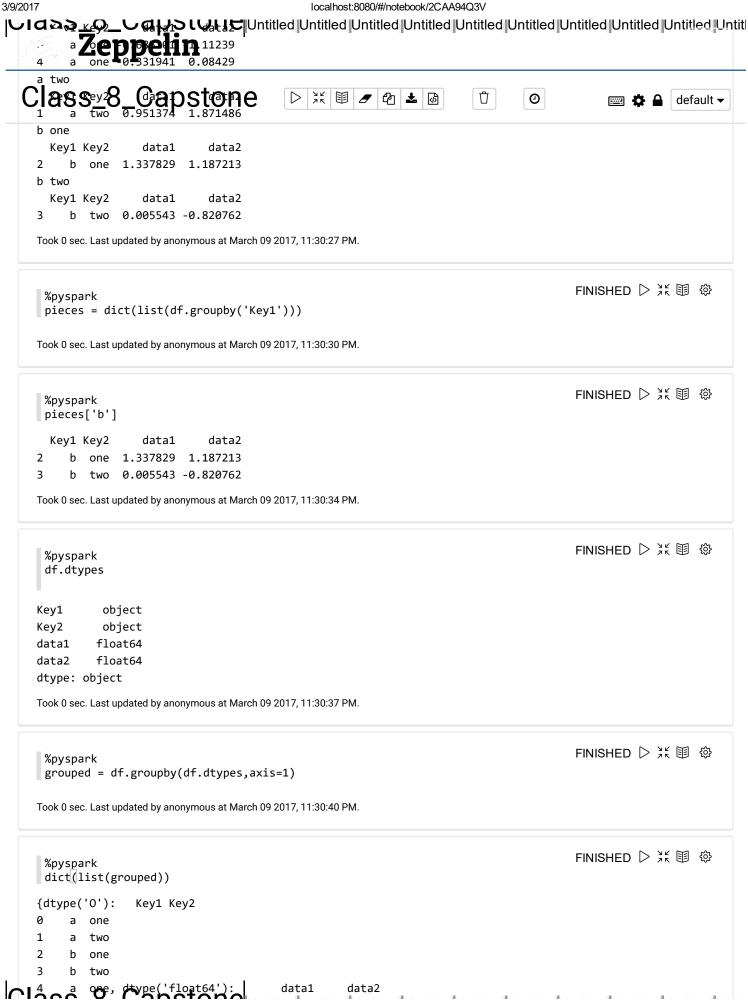
```
FINISHED ▷ 光 圓 ۞
 %pyspark
 df.groupby(['Key1','Key2']).mean()
               data1
                          data2
Key1 Key2
     one
          -0.508521 -0.514050
     two
            0.951374 1.871486
           1.337829 1.187213
h
     one
           0.005543 -0.820762
     two
Took 0 sec. Last updated by anonymous at March 09 2017, 11:30:16 PM.
```

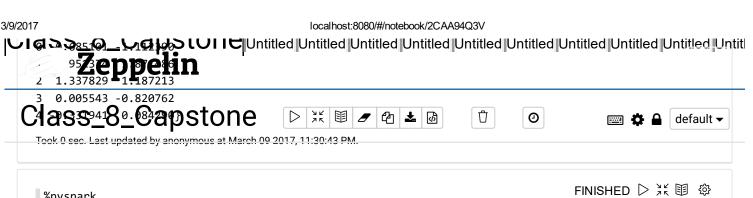
```
FINISHED ▷ ※ 圓 �
 %pyspark
 for name,group in df.groupby('Key1'):
     print name
     print group
а
  Key1 Key2
                 data1
                            data2
     a one -0.685101 -1.112390
1
              0.951374 1.871486
4
        one -0.331941
                        0.084290
b
  Key1 Key2
                 data1
                            data2
             1.337829 1.187213
2
        one
3
              0.005543 -0.820762
     b
        two
Took 0 sec. Last updated by anonymous at March 09 2017, 11:30:23 PM.
```

```
%pyspark
for (K1,K2),group in df.groupby(['Key1','Key2']):
    print K1,K2
    print group
```

FINISHED ▷ 茉 圓 �

Classe Q Canatana





```
%pyspark
 df.groupby('Key1')['data1']
<pandas.core.groupby.SeriesGroupBy object at 0x00000000006087630>
Took 0 sec. Last updated by anonymous at March 09 2017, 11:31:00 PM.
                                                                                          FINISHED ▷ 光 圓 墩
 %pyspark
df.groupby('Key1')[['data2']]
<pandas.core.groupby.DataFrameGroupBy object at 0x000000000629F470>
Took 0 sec. Last updated by anonymous at March 09 2017, 11:31:27 PM.
                                                                                          FINISHED ▷ 光 圓 ��
 %pyspark
df['data1'].groupby(df['Key1'])
<pandas.core.groupby.SeriesGroupBy object at 0x0000000006087668>
Took 0 sec. Last updated by anonymous at March 09 2017, 11:31:51 PM.
```

```
%pyspark
df[['data2']].groupby(df['Key1'])
```

<pandas.core.groupby.DataFrameGroupBy object at 0x0000000000060876D8> Took 0 sec. Last updated by anonymous at March 09 2017, 11:32:14 PM.

```
FINISHED ▷ ※ 圓 ��
 %pyspark
 df.groupby(['Key1', 'Key2'])[['data2']].mean()
               data2
Key1 Key2
          -0.514050
     one
            1.871486
     two
b
     one
            1.187213
          -0.820762
     two
Took 0 sec. Last updated by anonymous at March 09 2017, 11:32:47 PM.
```

```
FINISHED ▷ ※ 圓 ��
 %pyspark
 s_grouped = df.groupby(['Key1', 'Key2'])['data2']
s_grouped
<pandas.core.groupby.SeriesGroupBy object at 0x0000000003E876A0>
Took 0 sec. Last updated by anonymous at March 09 2017, 11:33:15 PM.
```

UIOSS\_O\_CAPSIUITE Untitled Unt

```
Key1 Key2
              _⁰©apstone
                                            紫 🎚 🗷 🖒 🕹
                                                                        Û
                                                                                ②
                                                                                                        default -
               1.187213
b
      one
              -0.820762
      two
Name: data2, dtype: float64
Took 0 sec. Last updated by anonymous at March 09 2017, 11:33:42 PM.
                                                                                        FINISHED ▷ 光 圓 ۞
 %pyspark
 people = pd.DataFrame(np.random.randn(5, 5),
  columns=['a', 'b', 'c', 'd', 'e'],
index=['Joe', 'Steve', 'Wes', 'Jim', 'Travis'])
Took 0 sec. Last updated by anonymous at March 09 2017, 11:34:21 PM.
                                                                                        FINISHED ▷ ※ 圓 ��
 %pyspark
 people.ix[2:3, ['b', 'c']] = np.nan # Add a few NA values
people
                          b
                                     C
                                                d
                                                           e
Joe
        0.423312 1.651784 1.658499 -1.604033 1.900242
Steve
        2.040601 1.454562 0.555265 -1.526755 -0.405573
       -0.204819
                        NaN
                                   NaN 1.343713 1.750284
Wes
Jim
       -0.427145   0.800888   -0.462763   0.597108   1.080201
Travis 0.200884 -0.387432 0.697745 0.516584 -1.654361
Took 0 sec. Last updated by anonymous at March 09 2017, 11:35:42 PM.
                                                                                        FINISHED ▷ ※ 圓 ��
 %pyspark
 mapping = {'a': 'red', 'b': 'red', 'c': 'blue',
  'd': 'blue', 'e': 'red', 'f' : 'orange'}
 by column = people.groupby(mapping, axis=1)
 by column.sum()
            blue
                        red
        0.054466 3.975339
Joe
Steve -0.971490 3.089590
        1.343713 1.545466
Wes
        0.134345 1.453945
Jim
Travis 1.214329 -1.840909
Took 0 sec. Last updated by anonymous at March 09 2017, 11:36:27 PM.
                                                                                       FINISHED ▷ 光 圓 ۞
 %pyspark
 map_series = pd.Series(mapping)
map_series
        red
        red
```

1500 pjæ

## UIGSS LO UGUSTUIT Untitled Unt

dtype: object 8 da Capstone 09 2017, 11:35:0 FM - 4 + 1 Û **②** default -FINISHED ▷ 兆 圓 ۞ %pyspark people.groupby(map series, axis=1).count() blue red 2 3 Joe Steve 2 3 2 Wes 1 Jim 2 3 Travis 2 3 Took 0 sec. Last updated by anonymous at March 09 2017, 11:37:46 PM. FINISHED ▷ ¾ ■ ∯ %pyspark people.groupby(len).sum() b c 3 -0.208652 2.452673 1.195736 0.336788 4.730728 2.040601 1.454562 0.555265 -1.526755 -0.405573 0.200884 -0.387432 0.697745 0.516584 -1.654361 Took 0 sec. Last updated by anonymous at March 09 2017, 11:38:26 PM. FINISHED ▷ 光 圓 �� %pyspark key\_list = ['one', 'one', 'one', 'two', 'two'] people.groupby([len, key\_list]).min() b d 3 one -0.204819 1.651784 1.658499 -1.604033 1.750284 two -0.427145 0.800888 -0.462763 0.597108 1.080201 5 one 2.040601 1.454562 0.555265 -1.526755 -0.405573 6 two 0.200884 -0.387432 0.697745 0.516584 -1.654361 Took 0 sec. Last updated by anonymous at March 09 2017, 11:39:20 PM. FINISHED ▷ 牂 圓 �� %pyspark columns = pd.MultiIndex.from\_arrays([['US', 'US', 'US', 'JP', 'JP'], [1, 3, 5, 1, 3]], names=['cty', 'tenor']) Took 0 sec. Last updated by anonymous at March 09 2017, 11:39:55 PM. FINISHED ▷ ※ 圓 �� %pyspark hier\_df = pd.DataFrame(np.random.randn(4, 5), columns=columns) hier\_df US JΡ cty 1 1 tenor 3

0.608612 -0.527968 1.540790 0.648186 -1.020264 0.000860 -1.005398 -0.698190 -1.419141 0.540718

0

3/9/2017 localhost:8080/#/notebook/2CAA94Q3V 

 0.063638 -3.539274
 0.333320 -0.109230
 0.823649

 Zeppenny -0.028447 -1.355646
 0.506471

 2 Took 0 sec. Last updated by anonymous at March 09 2017, 11:40:37 PM. Class\_8\_Capstone Û **②**  default 
 default %pyspark hier\_df.groupby(level='cty', axis=1).count() JP US 2 3 0 1 2 3 2 2 3 2 3 3 Took 0 sec. Last updated by anonymous at March 09 2017, 11:41:22 PM. READY ▷ 牂 圓 ��