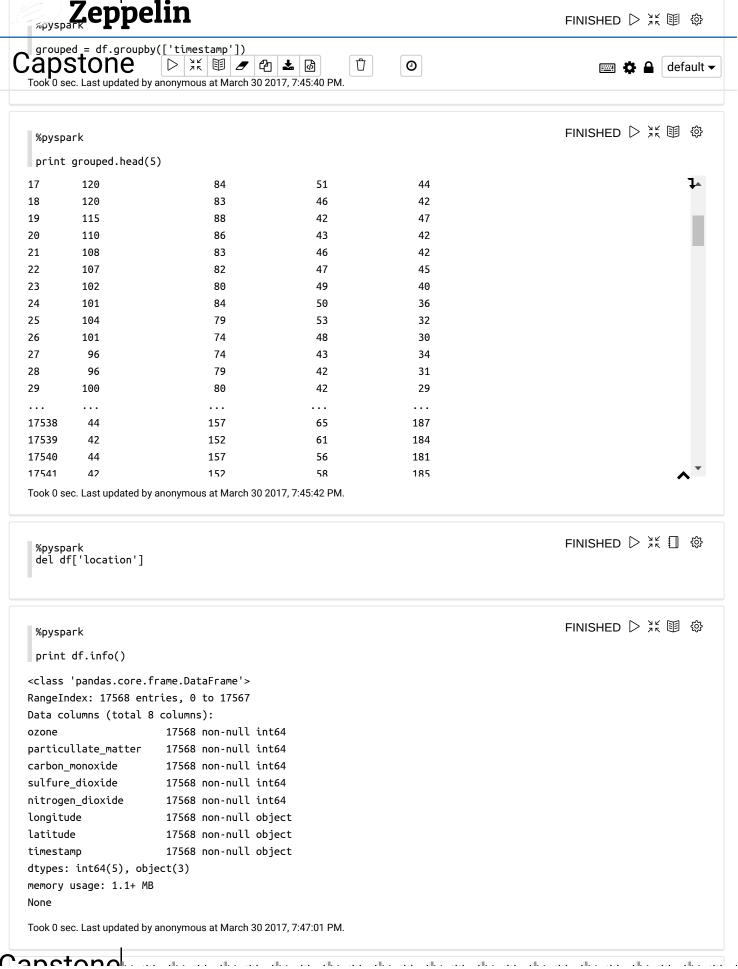
Zeppelin

```
Capstone
                                                            Ů
                                   ②
                                                                                                                   default -
                                                                                                 FINISHED ▷ ※ □ �
   %pyspark
    import os
    import glob
    from pandas import Series, DataFrame
    import pandas as pd
    import numpy as np
   p = '/home/gokul/Documents/pollution/'
   files = glob.glob(os.path.join(p, "*.csv"))
       = (pd.read csv(f) for f in files)
        = pd.concat(df, ignore_index=True)
                                                                                                 FINISHED ▷ 光 圓 ۞
   %pyspark
   df = pd.read_csv('/home/gokul/Documents/pollution/pollutionData158324.csv')
  Took 0 sec. Last updated by anonymous at March 30 2017, 7:33:07 PM.
                                                                                                 FINISHED ▷ 💥 🗐 🕸
   %pyspark
   df['longitude'] = df['longitude'].astype(str)
df['latitude'] = df['latitude'].astype(str)
   df["location"] = df[["longitude" ,"latitude"]].apply(lambda x: ','.join(x), axis=1)
  Took 1 sec. Last updated by anonymous at March 30 2017, 7:33:11 PM.
                                                                                                 FINISHED ▷ 端 圓 繳
   %pyspark
   print df.head(5)
             particullate_matter
     ozone
                                   carbon_monoxide
                                                     sulture_dloxide
  0
       101
                               94
                                                 49
                                                                   44
  1
       106
                               97
                                                 48
                                                                   47
  2
       107
                               95
                                                 49
                                                                   42
  3
       103
                               90
                                                 51
                                                                   44
  4
                               94
                                                 49
                                                                   39
       105
     nitrogen_dioxide
                             longitude
                                              latitude
                                                                   timestamp
  0
                        10.1049860761
                                        56.2317206943
                                                        2014-08-01 00:05:00
  1
                        10.1049860761
                                        56.2317206943
                                                        2014-08-01 00:10:00
  2
                        10.1049860761
                                        56.2317206943
                                                        2014-08-01 00:15:00
  3
                        10.1049860761
                                        56.2317206943
                                                        2014-08-01 00:20:00
  4
                        10.1049860761 56.2317206943 2014-08-01 00:25:00
                         location
     10.1049860761,56.2317206943
  0
     10.1049860761,56.2317206943
     10.1049860761,56.2317206943
     10.1049860761,56.2317206943
     10.1049860761,56.2317206943
  Took 0 sec. Last updated by anonymous at March 30 2017, 7:33:13 PM.
```

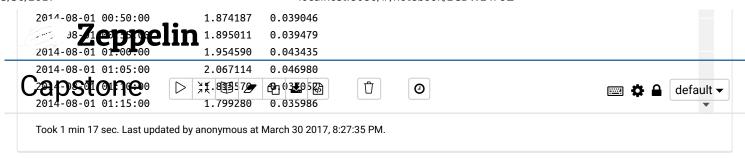
المرور والمراجع والمر

Untitled Unt



Untitled Unt

```
ozone_corr = lambda x: x.corrwith(['ozone'])
】②DiStOffepby(lookyx囯 tonestork)劢
                                                        Û
                                                                ②
                                                                                                             default -
  end = timeit.timeit()
 print end - start
Took 0 sec. Last updated by anonymous at March 30 2017, 7:54:43 PM. (outdated)
                                                                                              ERROR ▷ 湍 圓 墩
 %pyspark
 grouped.apply(df['particullate_matter'].corr(df['carbon_monoxide']))
Traceback (most recent call last):
  File "/tmp/zeppelin_pyspark-979765850639379021.py", line 267, in <module>
    raise Exception(traceback.format_exc())
Exception: Traceback (most recent call last):
  File "/tmp/zeppelin_pyspark-979765850639379021.py", line 265, in <module>
    exec(code)
  File "<stdin>", line 1, in <module>
  File "/home/gokul/anaconda2/lib/python2.7/site-packages/pandas/core/groupby.py", line 694, in apply
    return self._python_apply_general(f)
  File "/home/gokul/anaconda2/lib/python2.7/site-packages/pandas/core/groupby.py", line 698, in _python_apply_ge
neral
    self.axis)
  File "/home/gokul/anaconda2/lib/python2.7/site-packages/pandas/core/groupby.py", line 1611, in apply
    res = f(group)
TypeError: 'numpy.float64' object is not callable
Took 0 sec. Last updated by anonymous at March 30 2017, 8:10:01 PM. (outdated)
                                                                                            FINISHED ▷ ※ 圓 �
 %pyspark
  import timeit
  start = timeit.timeit()
  import statsmodels.api as sm
  def regression(data, yvar, xvars):
    Y = data[yvar]
    X = data[xvars]
    X['intercept'] = 1.
    result = sm.OLS(Y,X).fit()
    return result.params
  grouped.apply(regression, 'particullate_matter',['carbon_monoxide'])
                      carbon_monoxide intercept
timestamp
2014-08-01 00:05:00
                             1.917569
                                        0.039134
2014-08-01 00:10:00
                             2.019957
                                        0.042082
2014-08-01 00:15:00
                             1.937968
                                        0.039550
2014-08-01 00:20:00
                             1.764028
                                        0.034589
2014-08-01 00:25:00
                             1.917569
                                        0.039134
2014-08-01 00:30:00
                             1.915835
                                        0.039913
2014-08-01 00:35:00
                             1.739304
                                        0.034786
2014-08-01 00:40:00
                             1.749353
                                        0.033641
2014-08-01 00:45:00
                                        0.035186
                             1.759296
```



FINISHED D 洗 国 🕸

end = timeit.timeit()
print "time"
print end - start

time

%pyspark

-0.00243186950684

Took 1 min 14 sec. Last updated by anonymous at March 30 2017, 8:27:35 PM.

READY ▷ ※ 圓 ��