Capstone Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled

Zeppelin

Capstone

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"))

df  = (pd.read_csv(f) for f in files)
df   = pd.concat(df, ignore_index=True)
```

FINISHED

```
%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

```
%pyspark

print df.head(5)
```

```
  ozone  particullate_matter  carbon_monoxide  sulfure_dioxide  \
0   101                   94               49               44
1   106                   97               48               47
2   107                   95               49               42
3   103                   90               51               44
4   105                   94               49               39
  nitrogen_dioxide      longitude       latitude            timestamp  \
0               87  10.1049860761  56.2317206943  2014-08-01 00:05:00
1               86  10.1049860761  56.2317206943  2014-08-01 00:10:00
2               85  10.1049860761  56.2317206943  2014-08-01 00:15:00
3               87  10.1049860761  56.2317206943  2014-08-01 00:20:00
4               82  10.1049860761  56.2317206943  2014-08-01 00:25:00
                     location
0  10.1049860761,56.2317206943
1  10.1049860761,56.2317206943
2  10.1049860761,56.2317206943
3  10.1049860761,56.2317206943
4  10.1049860761,56.2317206943
```

Took 0 sec. Last updated by anonymous at March 30 2017, 7:33:13 PM.

Capstone

FINISHED

```
%pyspark

grouped = df.groupby(['timestamp'])
```

Took 0 sec. Last updated by anonymous at March 30 2017, 7:45:40 PM.

FINISHED

```
%pyspark

print grouped.head(5)
```

```
17       120              84              51              44
18       120              83              46              42
19       115              88              42              47
20       110              86              43              42
21       108              83              46              42
22       107              82              47              45
23       102              80              49              40
24       101              84              50              36
25       104              79              53              32
26       101              74              48              30
27        96              74              43              34
28        96              79              42              31
29       100              80              42              29
...      ...             ...             ...             ...
17538     44             157              65             187
17539     42             152              61             184
17540     44             157              56             181
17541     42             152              58             185
```

Took 0 sec. Last updated by anonymous at March 30 2017, 7:45:42 PM.

FINISHED

```
%pyspark
del df['location']
```

FINISHED

```
%pyspark

print df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17568 entries, 0 to 17567
Data columns (total 8 columns):
ozone                  17568 non-null int64
particullate_matter    17568 non-null int64
carbon_monoxide        17568 non-null int64
sulfure_dioxide        17568 non-null int64
nitrogen_dioxide       17568 non-null int64
longitude              17568 non-null object
latitude               17568 non-null object
timestamp              17568 non-null object
dtypes: int64(5), object(3)
memory usage: 1.1+ MB
None
```

Took 0 sec. Last updated by anonymous at March 30 2017, 7:47:01 PM.

FINISHED

```
%pyspark
import timeit
start = timeit.timeit()

ozone_corr = lambda x: x.corrwith(['ozone'])
by_time = df.groupby(lambda x: x['timestamp'])

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

Took 0 sec. Last updated by anonymous at March 30 2017, 7:54:43 PM. (outdated)

ERROR

```
%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

```
%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         1.759296   0.035186
```

```
2014-08-01 00:50:00         1.874187   0.039046
2014-08-01 00:55:00         1.895011   0.039479
2014-08-01 01:00:00         1.954590   0.043435
2014-08-01 01:05:00         2.067114   0.046980
2014-08-01 01:10:00         1.835570   0.037052
2014-08-01 01:15:00         1.799280   0.035986
```

Took 1 min 17 sec. Last updated by anonymous at March 30 2017, 8:27:35 PM.

FINISHED

```
%pyspark

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

```
time
-0.00243186950684
```

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

READY