May 11 - LAB 09

%pyspark FINISHED

```
import pandas import Series, DataFrame

import pandas as pd
```

Took 2 min 50 sec. Last updated by anonymous at May 11 2017, 8:25:12 PM.

%pyspark FINISHED

```
import numpy as np
```

Took 2 min 50 sec. Last updated by anonymous at May 11 2017, 8:27:29 PM.

```
%pysparkdf= DataFrame(

{'key1' : ['a', 'a', 'b', 'b', 'a'],'key2' : [ 'one', 'two',
'one', 'two', 'one'], 'data1' : np.random.randn(5),'data2' :
np.random.randn(5)})
```

```
data2 1.168887 1 -0.333507 -1.797469 2 -0.419739 -0.096406 3
0.586107 1.162645 4 -0.942160 -0.167812

key1 key2 a one a two b one b two a one
```

Took 2 min 50 sec. Last updated by anonymous at May 11 2017, 8:33:12 PM.

%pyspark FINISHED

```
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 2 min 50 sec. Last updated by anonymous at May 11 2017, 8:35:03 PM.

%pyspark FINISHED

```
print df.head(5)ozone accidents disturbance
```

%pyspark FINISHED

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

Took 3 sec. Last updated by anonymous at May 11 2017, 8:35:56 PM.

%pysparkdel df['location'] FINISHED

```
%pysparkprint df.info()

import timeitstart = timeit.timeit()
```

Took 6 sec. Last updated by anonymous at May 11 2017, 8:37:09 PM.

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

time 0.0048762098
```

Took 7 sec. Last updated by anonymous at May 11 2017, 8:39:27 PM.

%pyspark FINISHED

```
import timeitstart = timeit.timeit()

import statsmodels.api as smdef regression(data, yvar, xvars):

Y = data[yvar]X = data[xvars] X['intercept'] = 1. result =
sm.OLS(Y,X).fit() return result.params

grouped.apply(clustering,'burglaries,['latitude'])

int64 int64 int64 int64 int64 object object object
```

Took 7 sec. Last updated by anonymous at May 11 2017, 8:44:21 PM.