LabMar30 Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled

FINISHED

```
%pyspark
from pandas import Series, DataFrame
import pandas as pd
import numpy as np
```

Took 28 sec. Last updated by anonymous at March 30 2017, 6:53:42 PM.

FINISHED

```
%pyspark
data = DataFrame(pd.read_csv('/home/scarface/Desktop/sem-3/capstone/gsus.txt', sep=",", parse_dates=True, index_
```

Took 0 sec. Last updated by anonymous at March 30 2017, 8:07:41 PM.

FINISHED

```
%pyspark
#data=data[1:]
#data.drop(data.index[[0]], inplace=True)
data1=data["Open"]
data2=data["Close"]
data3=data["Low"]
#data1 = data[:data.shape[0]/2]
#data2 = data[data.shape[0]/2:]

dataF=DataFrame({'Open':data1,'Close':data2,'Low':data3})
```

Took 0 sec. Last updated by anonymous at March 30 2017, 8:21:53 PM.

FINISHED

```
%pyspark
dataF.info()
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 4487 entries, 1999-05-04 to 2017-03-07
Data columns (total 3 columns):
Close    4487 non-null float64
Low      4487 non-null float64
Open     4487 non-null float64
dtypes: float64(3)
memory usage: 140.2 KB
```

Took 0 sec. Last updated by anonymous at March 30 2017, 8:21:59 PM.

FINISHED

```
%pyspark
rets = dataF.pct_change().dropna()
```

Took 0 sec. Last updated by anonymous at March 30 2017, 8:22:10 PM.

FINISHED

```
%pyspark
high_corr = lambda x: x.corrwith(x['Low'])
by_year = rets.groupby(lambda x: x.year)
by_year.apply(high_corr)
```

	Close	Low	Open
1999	0.522126	1.0	0.745396
2000	0.668621	1.0	0.654431
2001	0.505661	1.0	0.677946
2002	0.612423	1.0	0.662827
2003	0.616051	1.0	0.654607
2004	0.620795	1.0	0.658016
2005	0.597846	1.0	0.677266
2006	0.670857	1.0	0.662410
2007	0.603077	1.0	0.708033
2008	0.756219	1.0	0.723269
2009	0.685850	1.0	0.655879
2010	0.681627	1.0	0.524637
2011	0.679694	1.0	0.676987
2012	0.646212	1.0	0.755057
2013	0.693622	1.0	0.664727
2014	0.742197	1.0	0.651536
2015	0.698222	1.0	0.782908

Took 0 sec. Last updated by anonymous at March 30 2017, 8:22:28 PM.

FINISHED

```
%pyspark
by_year.apply(lambda g: g['Open'].corr(g['Close']))
```

1999	0.228237
2000	0.154767
2001	0.[illegible]99395
2002	0.198649
2003	0.115[illegible]36
2004	0.202626
2005	0.147118
2006	0.185306
2007	0.225549
2008	0.390337
2009	0.134519
2010	0.078491
2011	0.276080
2012	0.243282
2013	0.229548
2014	0.302867
2015	0.315516
2016	0.461903

Took 0 sec. Last updated by anonymous at March 30 2017, 8:23:15 PM.

FINISHED

```
%pyspark
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
```

Took 3 sec. Last updated by anonymous at March 30 2017, 8:24:18 PM.

FINISHED

```
%pyspark
by_year.apply(regression,'Open',['Close'])
```

	Close	intercept
1999	0.209041	0.001044
2000	0.155496	0.001128
2001	0.093844	-0.000172
2002	0.186637	-0.000792
2003	0.111713	0.001453
2004	0.206454	0.000227
2005	0.149346	0.000701
2006	0.174542	0.001593
2007	0.234042	0.000329
2008	0.447949	-0.000990
2009	0.116849	0.002956
2010	0.087776	0.000226
2011	0.264422	-0.001567
2012	0.239452	0.001142
2013	0.223429	0.001166
2014	0.300502	0.000405
2015	0.305595	-0.000139

Took 1 sec. Last updated by anonymous at March 30 2017, 8:24:41 PM.

READY

LabMar30 Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled

Zeppelin

LabMar30

default