2-Pandas

September 3, 2016

1 Getting to know your data with Pandas

1.1 Pandas

Pandas is the Python Data Analysis Library.

Pandas is an extremely versatile tool for manipulating datasets.

It also produces high quality plots with matplotlib, and integrates nicely with other libraries that expect NumPy arrays.

The most important tool provided by Pandas is the **data frame.**

A data frame is a table in which each row and column is given a label.

Pandas DataFrames are documented at:
http://pandas.pydata.org/pandas-docs/dev/generated/pandas.DataFrame.html

1.2 Getting started

1.3 Fetching, storing and retrieving your data

For demonstration purposes, we'll use a library built-in to Pandas that fetches data from standard online sources, such as Yahoo! Finance.

More information on what types of data you can fetch is at: http://pandas.pydata.org/pandas-docs/stable/remote_data.html

```
In [267]: stocks = 'YELP'
          data_source = 'yahoo'
          start = datetime(2015,1,1)
          end = datetime(2015,12,31)

          yahoo_stocks = web.DataReader(stocks, data_source, start, end)

          # yahoo_stocks.head()
          yahoo_stocks.info()
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 252 entries, 2015-01-02 to 2015-12-31
Data columns (total 6 columns):
Open         252 non-null float64
High         252 non-null float64
Low          252 non-null float64
```

```
Close        252 non-null float64
Volume       252 non-null int64
Adj Close    252 non-null float64
dtypes: float64(5), int64(1)
memory usage: 13.8 KB
```

In [268]: yahoo_stocks

Out[268]:

Date	Open	High	Low	Close	Volume	Adj Clo
2015-01-02	55.459999	55.599998	54.240002	55.150002	1664500	55.1500
2015-01-05	54.540001	54.950001	52.330002	52.529999	2023000	52.5299
2015-01-06	52.549999	53.930000	50.750000	52.439999	3762800	52.4399
2015-01-07	53.320000	53.750000	51.759998	52.209999	1548200	52.2099
2015-01-08	52.590000	54.139999	51.759998	53.830002	2015300	53.8300
2015-01-09	55.959999	56.990002	54.720001	56.070000	6222600	56.0700
2015-01-12	56.000000	56.060001	53.430000	54.020000	2405100	54.0200
2015-01-13	54.470001	54.799999	52.520000	53.180000	1952100	53.1800
2015-01-14	52.799999	53.680000	51.459999	52.200001	1854600	52.2000
2015-01-15	53.000000	53.610001	50.029999	50.119999	2640400	50.1199
2015-01-16	50.180000	51.490002	50.029999	51.389999	2183300	51.3899
2015-01-20	51.650002	51.779999	50.689999	51.410000	1227600	51.4100
2015-01-21	51.200001	53.500000	51.200001	53.410000	3248100	53.4100
2015-01-22	53.869999	55.279999	53.119999	54.799999	2295400	54.7999
2015-01-23	54.660000	55.639999	54.299999	55.189999	1636400	55.1899
2015-01-26	55.119999	55.790001	54.830002	55.410000	1450300	55.4100
2015-01-27	56.060001	56.160000	54.570000	55.630001	2410400	55.6300
2015-01-28	56.150002	56.150002	52.919998	53.000000	2013100	53.0000
2015-01-29	52.849998	53.310001	51.410000	52.930000	1844100	52.9300
2015-01-30	52.590000	53.419998	52.049999	52.470001	1875400	52.4700
2015-02-02	52.939999	53.500000	51.209999	53.470001	2105500	53.4700
2015-02-03	53.830002	55.930000	53.410000	55.779999	2885400	55.7799
2015-02-04	55.529999	57.070000	55.250000	56.740002	2498600	56.7400
2015-02-05	57.599998	57.700001	56.080002	57.470001	4657300	57.4700
2015-02-06	47.700001	48.169998	44.860001	45.110001	25137400	45.1100
2015-02-09	44.910000	45.040001	42.099998	42.169998	13079300	42.1699
2015-02-10	43.830002	45.549999	43.310001	44.660000	11267700	44.6600
2015-02-11	45.389999	46.430000	44.810001	46.180000	6359400	46.1800
2015-02-12	46.450001	47.840000	45.950001	47.630001	4375000	47.6300
2015-02-13	48.509998	49.049999	47.220001	47.529999	4713100	47.5299
...	...	...	...	...	...	.
2015-11-18	27.540001	28.830000	27.309999	28.230000	3091600	28.2300
2015-11-19	28.190001	28.690001	27.910000	28.059999	1487500	28.0599
2015-11-20	28.100000	31.250000	28.049999	31.209999	6697500	31.2099
2015-11-23	30.580000	30.809999	29.150000	29.860001	4029900	29.8600
2015-11-24	29.459999	30.629999	29.450001	30.010000	2584500	30.0100
2015-11-25	29.790001	30.540001	29.709999	30.510000	1287100	30.5100

```
2015-11-27  30.500000  30.600000  29.610001  30.180000   1058900  30.1800
2015-11-30  30.110001  30.719999  29.770000  30.129999   2015600  30.1299
2015-12-01  30.110001  30.459999  29.799999  30.309999   1886000  30.3099
2015-12-02  30.299999  32.470001  30.290001  31.389999   4650300  31.3899
2015-12-03  31.389999  32.240002  30.480000  30.629999   2698900  30.6299
2015-12-04  30.530001  30.860001  29.320000  30.450001   2313800  30.4500
2015-12-07  30.379999  30.639999  29.629999  30.040001   1362300  30.0400
2015-12-08  29.809999  31.379999  29.500000  30.920000   1830200  30.9200
2015-12-09  30.980000  31.139999  29.260000  30.000000   2238500  30.0000
2015-12-10  30.110001  31.299999  29.990000  30.830000   1252900  30.8300
2015-12-11  30.690001  30.750000  29.600000  29.650000   1415000  29.6500
2015-12-14  29.600000  29.889999  28.850000  29.580000   2328600  29.5800
2015-12-15  29.680000  30.000000  26.459999  26.870001   5759200  26.8700
2015-12-16  26.889999  28.240000  26.260000  28.030001   2992100  28.0300
2015-12-17  28.139999  28.320000  27.190001  27.420000   1483900  27.4200
2015-12-18  27.309999  27.910000  26.900000  27.170000   1299800  27.1700
2015-12-21  27.170000  27.360001  26.030001  26.250000   1947600  26.2500
2015-12-22  26.250000  28.700001  26.150000  27.930000   2952700  27.9300
2015-12-23  27.950001  28.420000  27.440001  28.150000   1001000  28.1500
2015-12-24  28.270000  28.590000  27.900000  28.400000    587400  28.4000
2015-12-28  28.120001  28.379999  27.770000  27.879999   1004500  27.8799
2015-12-29  27.950001  28.540001  27.740000  28.480000   1103900  28.4800
2015-12-30  28.580000  28.780001  28.170000  28.250000   1068000  28.2500
2015-12-31  28.100000  28.969999  28.020000  28.799999   1301500  28.7999

[252 rows x 6 columns]
```

1.3.1 Reading data from a .csv file

```
In [269]: yahoo_stocks.to_csv('yahoo_data.csv')
          #print(open('yahoo_data.csv').read())

In [270]: df = pd.read_csv('yahoo_data.csv')
          df
```

Out[270]:

	Date	Open	High	Low	Close	Volume	\
0	2015-01-02	55.459999	55.599998	54.240002	55.150002	1664500	
1	2015-01-05	54.540001	54.950001	52.330002	52.529999	2023000	
2	2015-01-06	52.549999	53.930000	50.750000	52.439999	3762800	
3	2015-01-07	53.320000	53.750000	51.759998	52.209999	1548200	
4	2015-01-08	52.590000	54.139999	51.759998	53.830002	2015300	
5	2015-01-09	55.959999	56.990002	54.720001	56.070000	6222600	
6	2015-01-12	56.000000	56.060001	53.430000	54.020000	2405100	
7	2015-01-13	54.470001	54.799999	52.520000	53.180000	1952100	
8	2015-01-14	52.799999	53.680000	51.459999	52.200001	1854600	
9	2015-01-15	53.000000	53.610001	50.029999	50.119999	2640400	
10	2015-01-16	50.180000	51.490002	50.029999	51.389999	2183300	
11	2015-01-20	51.650002	51.779999	50.689999	51.410000	1227600	

12	2015-01-21	51.200001	53.500000	51.200001	53.410000	3248100
13	2015-01-22	53.869999	55.279999	53.119999	54.799999	2295400
14	2015-01-23	54.660000	55.639999	54.299999	55.189999	1636400
15	2015-01-26	55.119999	55.790001	54.830002	55.410000	1450300
16	2015-01-27	56.060001	56.160000	54.570000	55.630001	2410400
17	2015-01-28	56.150002	56.150002	52.919998	53.000000	2013100
18	2015-01-29	52.849998	53.310001	51.410000	52.930000	1844100
19	2015-01-30	52.590000	53.419998	52.049999	52.470001	1875400
20	2015-02-02	52.939999	53.500000	51.209999	53.470001	2105500
21	2015-02-03	53.830002	55.930000	53.410000	55.779999	2885400
22	2015-02-04	55.529999	57.070000	55.250000	56.740002	2498600
23	2015-02-05	57.599998	57.700001	56.080002	57.470001	4657300
24	2015-02-06	47.700001	48.169998	44.860001	45.110001	25137400
25	2015-02-09	44.910000	45.040001	42.099998	42.169998	13079300
26	2015-02-10	43.830002	45.549999	43.310001	44.660000	11267700
27	2015-02-11	45.389999	46.430000	44.810001	46.180000	6359400
28	2015-02-12	46.450001	47.840000	45.950001	47.630001	4375000
29	2015-02-13	48.509998	49.049999	47.220001	47.529999	4713100
..	...	...	...	...	...	...
222	2015-11-18	27.540001	28.830000	27.309999	28.230000	3091600
223	2015-11-19	28.190001	28.690001	27.910000	28.059999	1487500
224	2015-11-20	28.100000	31.250000	28.049999	31.209999	6697500
225	2015-11-23	30.580000	30.809999	29.150000	29.860001	4029900
226	2015-11-24	29.459999	30.629999	29.450001	30.010000	2584500
227	2015-11-25	29.790001	30.540001	29.709999	30.510000	1287100
228	2015-11-27	30.500000	30.600000	29.610001	30.180000	1058900
229	2015-11-30	30.110001	30.719999	29.770000	30.129999	2015600
230	2015-12-01	30.110001	30.459999	29.799999	30.309999	1886000
231	2015-12-02	30.299999	32.470001	30.290001	31.389999	4650300
232	2015-12-03	31.389999	32.240002	30.480000	30.629999	2698900
233	2015-12-04	30.530001	30.860001	29.320000	30.450001	2313800
234	2015-12-07	30.379999	30.639999	29.629999	30.040001	1362300
235	2015-12-08	29.809999	31.379999	29.500000	30.920000	1830200
236	2015-12-09	30.980000	31.139999	29.260000	30.000000	2238500
237	2015-12-10	30.110001	31.299999	29.990000	30.830000	1252900
238	2015-12-11	30.690001	30.750000	29.600000	29.650000	1415000
239	2015-12-14	29.600000	29.889999	28.850000	29.580000	2328600
240	2015-12-15	29.680000	30.000000	26.459999	26.870001	5759200
241	2015-12-16	26.889999	28.240000	26.260000	28.030001	2992100
242	2015-12-17	28.139999	28.320000	27.190001	27.420000	1483900
243	2015-12-18	27.309999	27.910000	26.900000	27.170000	1299800
244	2015-12-21	27.170000	27.360001	26.030001	26.250000	1947600
245	2015-12-22	26.250000	28.700001	26.150000	27.930000	2952700
246	2015-12-23	27.950001	28.420000	27.440001	28.150000	1001000
247	2015-12-24	28.270000	28.590000	27.900000	28.400000	587400
248	2015-12-28	28.120001	28.379999	27.770000	27.879999	1004500
249	2015-12-29	27.950001	28.540001	27.740000	28.480000	1103900
250	2015-12-30	28.580000	28.780001	28.170000	28.250000	1068000

```
251  2015-12-31  28.100000  28.969999  28.020000  28.799999   1301500
```

```
     Adj Close
0    55.150002
1    52.529999
2    52.439999
3    52.209999
4    53.830002
5    56.070000
6    54.020000
7    53.180000
8    52.200001
9    50.119999
10   51.389999
11   51.410000
12   53.410000
13   54.799999
14   55.189999
15   55.410000
16   55.630001
17   53.000000
18   52.930000
19   52.470001
20   53.470001
21   55.779999
22   56.740002
23   57.470001
24   45.110001
25   42.169998
26   44.660000
27   46.180000
28   47.630001
29   47.529999
..         ...
222  28.230000
223  28.059999
224  31.209999
225  29.860001
226  30.010000
227  30.510000
228  30.180000
229  30.129999
230  30.309999
231  31.389999
232  30.629999
233  30.450001
234  30.040001
235  30.920000
```

```
236  30.000000
237  30.830000
238  29.650000
239  29.580000
240  26.870001
241  28.030001
242  27.420000
243  27.170000
244  26.250000
245  27.930000
246  28.150000
247  28.400000
248  27.879999
249  28.480000
250  28.250000
251  28.799999

[252 rows x 7 columns]
```

The number of rows in the DataFrame:

```
In [271]: len(df)

Out[271]: 252
```

1.4 Working with data columns

The columns or "features" in your data

```
In [272]: df.columns

Out[272]: Index(['Date', 'Open', 'High', 'Low', 'Close', 'Volume', 'Adj Close'], dt
```

Selecting a single column from your data

```
In [273]: df['Open']

Out[273]: 0      55.459999
          1      54.540001
          2      52.549999
          3      53.320000
          4      52.590000
          5      55.959999
          6      56.000000
          7      54.470001
          8      52.799999
          9      53.000000
          10     50.180000
          11     51.650002
```

```
12     51.200001
13     53.869999
14     54.660000
15     55.119999
16     56.060001
17     56.150002
18     52.849998
19     52.590000
20     52.939999
21     53.830002
22     55.529999
23     57.599998
24     47.700001
25     44.910000
26     43.830002
27     45.389999
28     46.450001
29     48.509998
         ...
222    27.540001
223    28.190001
224    28.100000
225    30.580000
226    29.459999
227    29.790001
228    30.500000
229    30.110001
230    30.110001
231    30.299999
232    31.389999
233    30.530001
234    30.379999
235    29.809999
236    30.980000
237    30.110001
238    30.690001
239    29.600000
240    29.680000
241    26.889999
242    28.139999
243    27.309999
244    27.170000
245    26.250000
246    27.950001
247    28.270000
248    28.120001
249    27.950001
250    28.580000
```

```
251    28.100000
Name: Open, dtype: float64
```

Another way of selecting a single column from your data

```
In [274]: df.Open

Out[274]: 0      55.459999
          1      54.540001
          2      52.549999
          3      53.320000
          4      52.590000
          5      55.959999
          6      56.000000
          7      54.470001
          8      52.799999
          9      53.000000
          10     50.180000
          11     51.650002
          12     51.200001
          13     53.869999
          14     54.660000
          15     55.119999
          16     56.060001
          17     56.150002
          18     52.849998
          19     52.590000
          20     52.939999
          21     53.830002
          22     55.529999
          23     57.599998
          24     47.700001
          25     44.910000
          26     43.830002
          27     45.389999
          28     46.450001
          29     48.509998
                   ...
          222    27.540001
          223    28.190001
          224    28.100000
          225    30.580000
          226    29.459999
          227    29.790001
          228    30.500000
          229    30.110001
          230    30.110001
          231    30.299999
```

```
232    31.389999
233    30.530001
234    30.379999
235    29.809999
236    30.980000
237    30.110001
238    30.690001
239    29.600000
240    29.680000
241    26.889999
242    28.139999
243    27.309999
244    27.170000
245    26.250000
246    27.950001
247    28.270000
248    28.120001
249    27.950001
250    28.580000
251    28.100000
Name: Open, dtype: float64
```

```
In [275]: df[['Open','Close']].head()
```

```
Out[275]:         Open      Close
          0  55.459999  55.150002
          1  54.540001  52.529999
          2  52.549999  52.439999
          3  53.320000  52.209999
          4  52.590000  53.830002
```

```
In [276]: df.Date.head(10)
```

```
Out[276]: 0    2015-01-02
          1    2015-01-05
          2    2015-01-06
          3    2015-01-07
          4    2015-01-08
          5    2015-01-09
          6    2015-01-12
          7    2015-01-13
          8    2015-01-14
          9    2015-01-15
          Name: Date, dtype: object
```

```
In [277]: df.Date.tail(10)
```

```
Out[277]: 242    2015-12-17
          243    2015-12-18
```

```
244    2015-12-21
245    2015-12-22
246    2015-12-23
247    2015-12-24
248    2015-12-28
249    2015-12-29
250    2015-12-30
251    2015-12-31
Name: Date, dtype: object
```

Changing the column names:

```
In [278]: new_column_names = [x.lower().replace(' ','_') for x in df.columns]
          df.columns = new_column_names
          df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 252 entries, 0 to 251
Data columns (total 7 columns):
date         252 non-null object
open         252 non-null float64
high         252 non-null float64
low          252 non-null float64
close        252 non-null float64
volume       252 non-null int64
adj_close    252 non-null float64
dtypes: float64(5), int64(1), object(1)
memory usage: 13.9+ KB
```

Now **all** columns can be accessed using the **dot** notation:

```
In [279]: df.adj_close.head()
```

```
Out[279]: 0    55.150002
          1    52.529999
          2    52.439999
          3    52.209999
          4    53.830002
          Name: adj_close, dtype: float64
```

1.5 Data Frame methods

A DataFrame object has many useful methods.

```
In [280]: df.mean()
```

```
Out[280]: open         3.728766e+01
          high         3.805464e+01
```

```
low          3.656373e+01
close        3.729917e+01
volume       3.492134e+06
adj_close    3.729917e+01
dtype: float64
```

```
In [281]: df.std()
```

```
Out[281]: open         1.128093e+01
          high         1.138111e+01
          low          1.113097e+01
          close        1.125233e+01
          volume       4.145502e+06
          adj_close    1.125233e+01
          dtype: float64
```

```
In [282]: df.median()
```

```
Out[282]: open         3.796500e+01
          high         3.871500e+01
          low          3.637500e+01
          close        3.783500e+01
          volume       2.354050e+06
          adj_close    3.783500e+01
          dtype: float64
```

```
In [283]: df.open.mean()
```

```
Out[283]: 37.287658698412692
```

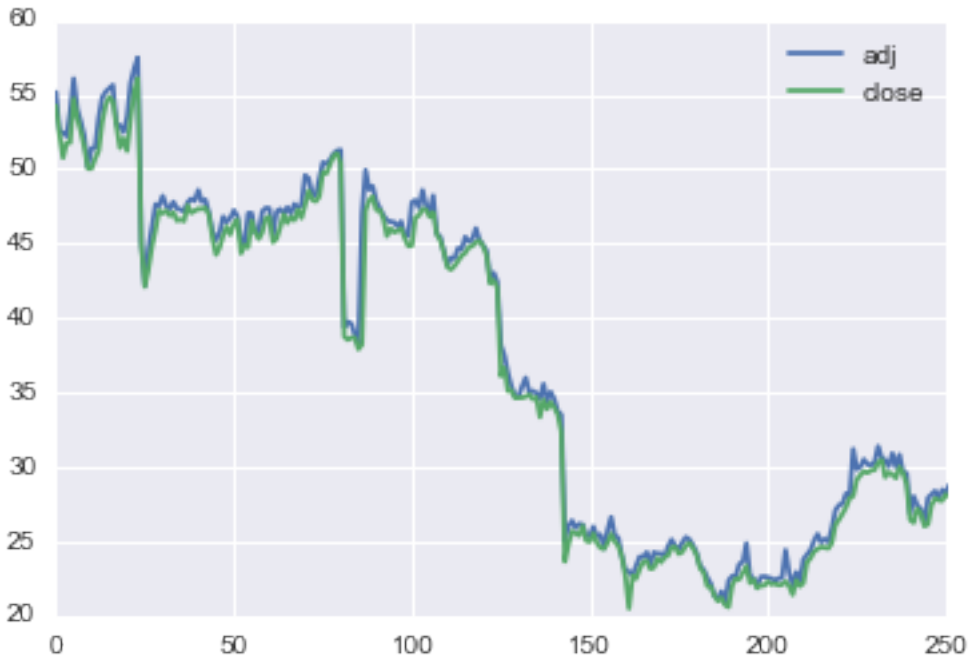
```
In [284]: df.high.mean()
```

```
Out[284]: 38.054642952380952
```

1.5.1 Plotting methods

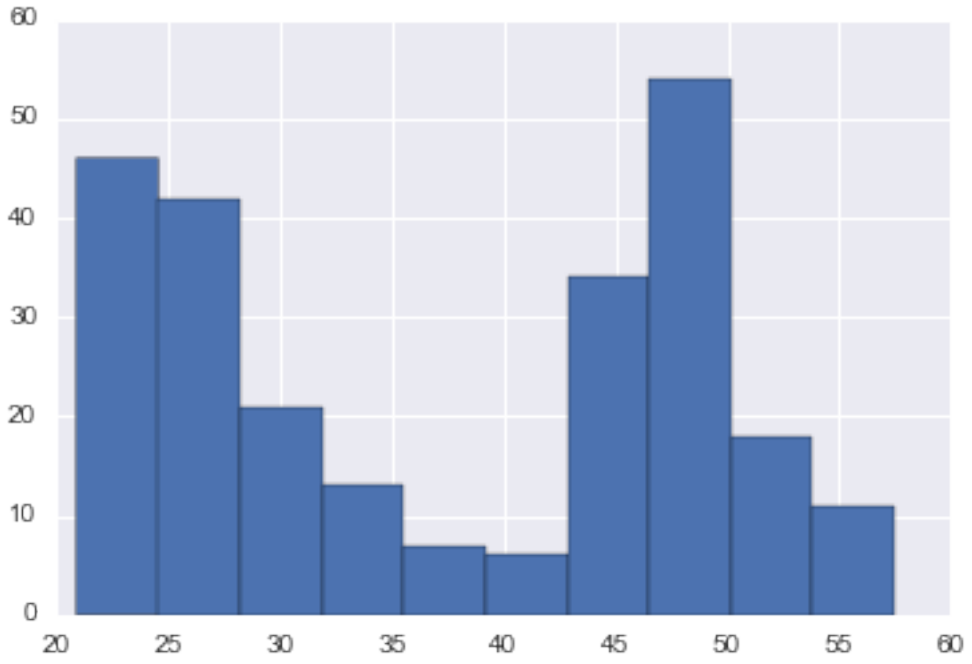
```
In [285]: df.adj_close.plot(label='adj')
          df.low.plot(label='close')
          plt.legend(loc='best')
```

```
Out[285]: <matplotlib.legend.Legend at 0x11e74c748>
```




```
In [286]: df.adj_close.hist()
```

```
Out[286]: <matplotlib.axes._subplots.AxesSubplot at 0x11f069f98>
```



1.5.2 Bulk Operations

Methods like **sum( )** and **std( )** work on entire columns.

We can run our own functions across all values in a column (or row) using **apply( )**.

```
In [287]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 252 entries, 0 to 251
Data columns (total 7 columns):
date         252 non-null object
open         252 non-null float64
high         252 non-null float64
low          252 non-null float64
close        252 non-null float64
volume       252 non-null int64
adj_close    252 non-null float64
dtypes: float64(5), int64(1), object(1)
memory usage: 13.9+ KB
```

```
In [288]: df.date.head()
```

```
Out[288]: 0    2015-01-02
          1    2015-01-05
          2    2015-01-06
          3    2015-01-07
          4    2015-01-08
          Name: date, dtype: object
```

The **values** property of the column returns a list of values for the column. Inspecting the first value reveals that these are strings with a particular format.

```
In [289]: first_date = df.date.values[0]
          first_date
```

```
Out[289]: '2015-01-02'
```

```
In [290]: datetime.strptime(first_date, "%Y-%m-%d")
```

```
Out[290]: datetime.datetime(2015, 1, 2, 0, 0)
```

```
In [291]: df.date = df.date.apply(lambda d: datetime.strptime(d, "%Y-%m-%d"))
          df.date.head()
```

```
Out[291]: 0   2015-01-02
          1   2015-01-05
          2   2015-01-06
          3   2015-01-07
          4   2015-01-08
          Name: date, dtype: datetime64[ns]
```

Each row in a DataFrame is associated with an index, which is a label that uniquely identifies a row.

The row indices so far have been auto-generated by pandas, and are simply integers starting from 0.

From now on we will use dates instead of integers for indices – the benefits of this will show later.

Overwriting the index is as easy as assigning to the **`index`** property of the DataFrame.

```
In [292]: df.index = df.date
          df.head()
```

```
Out[292]:                  date       open       high        low      close   volum
          date
          2015-01-02 2015-01-02  55.459999  55.599998  54.240002  55.150002  166450
          2015-01-05 2015-01-05  54.540001  54.950001  52.330002  52.529999  202300
          2015-01-06 2015-01-06  52.549999  53.930000  50.750000  52.439999  376280
          2015-01-07 2015-01-07  53.320000  53.750000  51.759998  52.209999  154820
          2015-01-08 2015-01-08  52.590000  54.139999  51.759998  53.830002  201530

                      adj_close
          date
          2015-01-02  55.150002
          2015-01-05  52.529999
          2015-01-06  52.439999
          2015-01-07  52.209999
          2015-01-08  53.830002
```

Now that we have made an index based on date, we can drop the original `date` column.

```
In [293]: df = df.drop(['date'],axis=1)
          df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 252 entries, 2015-01-02 to 2015-12-31
Data columns (total 6 columns):
open         252 non-null float64
high         252 non-null float64
low          252 non-null float64
close        252 non-null float64
volume       252 non-null int64
adj_close    252 non-null float64
dtypes: float64(5), int64(1)
memory usage: 13.8 KB
```

1.5.3 Accessing rows of the DataFrame

So far we've seen how to access a column of the DataFrame. To access a row we use a different notation.

To access a row by its index value, use the **`.ix()`** method.

```
In [294]: df.ix[datetime(2015,1,23,0,0)]

Out[294]: open         5.466000e+01
          high         5.564000e+01
          low          5.430000e+01
          close        5.519000e+01
          volume       1.636400e+06
          adj_close    5.519000e+01
          Name: 2015-01-23 00:00:00, dtype: float64
```

To access a row by its sequence number (ie, like an array index), use **`.iloc()`** ('Integer Location')

```
In [295]: df.iloc[0,:]

Out[295]: open         5.546000e+01
          high         5.560000e+01
          low          5.424000e+01
          close        5.515000e+01
          volume       1.664500e+06
          adj_close    5.515000e+01
          Name: 2015-01-02 00:00:00, dtype: float64
```

To iterate over the rows, use **`.iterrows()`**

```
In [296]: num_positive_days = 0
          for idx, row in df.iterrows():
              if row.close > row.open:
                  num_positive_days += 1

          print("The total number of positive-gain days is {}.".format(num_positive

The total number of positive-gain days is 126.
```

1.6 Filtering

It is very easy to select interesting rows from the data.

All these operations below return a new DataFrame, which itself can be treated the same way as all DataFrames we have seen so far.

```
In [297]: tmp_high = df.high > 55
          tmp_high.head()

Out[297]: date
          2015-01-02     True
          2015-01-05    False
          2015-01-06    False
          2015-01-07    False
          2015-01-08    False
          Name: high, dtype: bool
```

Summing a Boolean array is the same as counting the number of **`True`** values.

```
In [298]: sum(tmp_high)

Out[298]: 11
```

Now, let's select only the rows of **`df1`** that correspond to **`tmp_high`**

```
In [299]: df[tmp_high]

Out[299]:                  open       high        low      close   volume  adj_clos
          date
          2015-01-02  55.459999  55.599998  54.240002  55.150002  1664500  55.15000
          2015-01-09  55.959999  56.990002  54.720001  56.070000  6222600  56.07000
          2015-01-12  56.000000  56.060001  53.430000  54.020000  2405100  54.02000
          2015-01-22  53.869999  55.279999  53.119999  54.799999  2295400  54.79999
          2015-01-23  54.660000  55.639999  54.299999  55.189999  1636400  55.18999
          2015-01-26  55.119999  55.790001  54.830002  55.410000  1450300  55.41000
          2015-01-27  56.060001  56.160000  54.570000  55.630001  2410400  55.63000
          2015-01-28  56.150002  56.150002  52.919998  53.000000  2013100  53.00000
          2015-02-03  53.830002  55.930000  53.410000  55.779999  2885400  55.77999
          2015-02-04  55.529999  57.070000  55.250000  56.740002  2498600  56.74000
          2015-02-05  57.599998  57.700001  56.080002  57.470001  4657300  57.47000
```

Putting it all together, we have the following commonly-used patterns:

```
In [300]: positive_days = df[df.close > df.open]
          positive_days.head()

Out[300]:                  open       high        low      close   volume  adj_clos
          date
          2015-01-08  52.590000  54.139999  51.759998  53.830002  2015300  53.83000
          2015-01-09  55.959999  56.990002  54.720001  56.070000  6222600  56.07000
          2015-01-16  50.180000  51.490002  50.029999  51.389999  2183300  51.38999
          2015-01-21  51.200001  53.500000  51.200001  53.410000  3248100  53.41000
          2015-01-22  53.869999  55.279999  53.119999  54.799999  2295400  54.79999

In [301]: very_positive_days = df[df.close-df.open > 4]
          very_positive_days.head()

Out[301]:                  open   high        low      close    volume  adj_close
          date
          2015-05-07  38.220001  48.73  38.220001  47.009998  33831600  47.009998
```

1.7 Creating new columns

To create a new column, simply assign values to it. Think of the columns as a dictionary:

```
In [302]: df['profit'] = (df.open < df.close)
          df.head()
```

```
Out[302]:                 open       high        low      close   volume  adj_clos
          date
          2015-01-02  55.459999  55.599998  54.240002  55.150002  1664500  55.15000
          2015-01-05  54.540001  54.950001  52.330002  52.529999  2023000  52.52999
          2015-01-06  52.549999  53.930000  50.750000  52.439999  3762800  52.43999
          2015-01-07  53.320000  53.750000  51.759998  52.209999  1548200  52.20999
          2015-01-08  52.590000  54.139999  51.759998  53.830002  2015300  53.83000

                      profit
          date
          2015-01-02   False
          2015-01-05   False
          2015-01-06   False
          2015-01-07   False
          2015-01-08    True
```

```
In [303]: for idx, row in df.iterrows():
              if row.close > row.open:
                  df.ix[idx,'gain']='negative'
              elif (row.open - row.close) < 1:
                  df.ix[idx,'gain']='small_gain'
              elif (row.open - row.close) < 6:
                  df.ix[idx,'gain']='medium_gain'
              else:
                  df.ix[idx,'gain']='large_gain'
          df.head()
```

```
Out[303]:                 open       high        low      close   volume  adj_clos
          date
          2015-01-02  55.459999  55.599998  54.240002  55.150002  1664500  55.15000
          2015-01-05  54.540001  54.950001  52.330002  52.529999  2023000  52.52999
          2015-01-06  52.549999  53.930000  50.750000  52.439999  3762800  52.43999
          2015-01-07  53.320000  53.750000  51.759998  52.209999  1548200  52.20999
          2015-01-08  52.590000  54.139999  51.759998  53.830002  2015300  53.83000

                      profit         gain
          date
          2015-01-02   False   small_gain
          2015-01-05   False  medium_gain
          2015-01-06   False   small_gain
          2015-01-07   False  medium_gain
          2015-01-08    True     negative
```

Here is another, more "functional", way to accomplish the same thing. Define a function that classifies rows, and **`apply`** it to each row.

```
In [304]: def namerow(row):
              if row.close > row.open:
                  return 'negative'
```

```
            elif (row.open - row.close) < 1:
                return 'small_gain'
            elif (row.open - row.close) < 6:
                return 'medium_gain'
            else:
                return 'large_gain'

        df['test_column'] = df.apply(namerow, axis = 1)
```

```
In [305]: df.head()
```

Out[305]:

date	open	high	low	close	volume	adj_clos
2015-01-02	55.459999	55.599998	54.240002	55.150002	1664500	55.15000
2015-01-05	54.540001	54.950001	52.330002	52.529999	2023000	52.52999
2015-01-06	52.549999	53.930000	50.750000	52.439999	3762800	52.43999
2015-01-07	53.320000	53.750000	51.759998	52.209999	1548200	52.20999
2015-01-08	52.590000	54.139999	51.759998	53.830002	2015300	53.83000

date	profit	gain	test_column
2015-01-02	False	small_gain	small_gain
2015-01-05	False	medium_gain	medium_gain
2015-01-06	False	small_gain	small_gain
2015-01-07	False	medium_gain	medium_gain
2015-01-08	True	negative	negative

OK, point made, let's get rid of that extraneous `test_column`:

```
In [306]: df.drop('test_column', axis = 1)
```

Out[306]:

date	open	high	low	close	volume	adj_clo
2015-01-02	55.459999	55.599998	54.240002	55.150002	1664500	55.1500
2015-01-05	54.540001	54.950001	52.330002	52.529999	2023000	52.5299
2015-01-06	52.549999	53.930000	50.750000	52.439999	3762800	52.4399
2015-01-07	53.320000	53.750000	51.759998	52.209999	1548200	52.2099
2015-01-08	52.590000	54.139999	51.759998	53.830002	2015300	53.8300
2015-01-09	55.959999	56.990002	54.720001	56.070000	6222600	56.0700
2015-01-12	56.000000	56.060001	53.430000	54.020000	2405100	54.0200
2015-01-13	54.470001	54.799999	52.520000	53.180000	1952100	53.1800
2015-01-14	52.799999	53.680000	51.459999	52.200001	1854600	52.2000
2015-01-15	53.000000	53.610001	50.029999	50.119999	2640400	50.1199
2015-01-16	50.180000	51.490002	50.029999	51.389999	2183300	51.3899
2015-01-20	51.650002	51.779999	50.689999	51.410000	1227600	51.4100
2015-01-21	51.200001	53.500000	51.200001	53.410000	3248100	53.4100
2015-01-22	53.869999	55.279999	53.119999	54.799999	2295400	54.7999
2015-01-23	54.660000	55.639999	54.299999	55.189999	1636400	55.1899
2015-01-26	55.119999	55.790001	54.830002	55.410000	1450300	55.4100

2015-01-27	56.060001	56.160000	54.570000	55.630001	2410400	55.6300
2015-01-28	56.150002	56.150002	52.919998	53.000000	2013100	53.0000
2015-01-29	52.849998	53.310001	51.410000	52.930000	1844100	52.9300
2015-01-30	52.590000	53.419998	52.049999	52.470001	1875400	52.4700
2015-02-02	52.939999	53.500000	51.209999	53.470001	2105500	53.4700
2015-02-03	53.830002	55.930000	53.410000	55.779999	2885400	55.7799
2015-02-04	55.529999	57.070000	55.250000	56.740002	2498600	56.7400
2015-02-05	57.599998	57.700001	56.080002	57.470001	4657300	57.4700
2015-02-06	47.700001	48.169998	44.860001	45.110001	25137400	45.1100
2015-02-09	44.910000	45.040001	42.099998	42.169998	13079300	42.1699
2015-02-10	43.830002	45.549999	43.310001	44.660000	11267700	44.6600
2015-02-11	45.389999	46.430000	44.810001	46.180000	6359400	46.1800
2015-02-12	46.450001	47.840000	45.950001	47.630001	4375000	47.6300
2015-02-13	48.509998	49.049999	47.220001	47.529999	4713100	47.5299
...	...	...	...	...	...	.
2015-11-18	27.540001	28.830000	27.309999	28.230000	3091600	28.2300
2015-11-19	28.190001	28.690001	27.910000	28.059999	1487500	28.0599
2015-11-20	28.100000	31.250000	28.049999	31.209999	6697500	31.2099
2015-11-23	30.580000	30.809999	29.150000	29.860001	4029900	29.8600
2015-11-24	29.459999	30.629999	29.450001	30.010000	2584500	30.0100
2015-11-25	29.790001	30.540001	29.709999	30.510000	1287100	30.5100
2015-11-27	30.500000	30.600000	29.610001	30.180000	1058900	30.1800
2015-11-30	30.110001	30.719999	29.770000	30.129999	2015600	30.1299
2015-12-01	30.110001	30.459999	29.799999	30.309999	1886000	30.3099
2015-12-02	30.299999	32.470001	30.290001	31.389999	4650300	31.3899
2015-12-03	31.389999	32.240002	30.480000	30.629999	2698900	30.6299
2015-12-04	30.530001	30.860001	29.320000	30.450001	2313800	30.4500
2015-12-07	30.379999	30.639999	29.629999	30.040001	1362300	30.0400
2015-12-08	29.809999	31.379999	29.500000	30.920000	1830200	30.9200
2015-12-09	30.980000	31.139999	29.260000	30.000000	2238500	30.0000
2015-12-10	30.110001	31.299999	29.990000	30.830000	1252900	30.8300
2015-12-11	30.690001	30.750000	29.600000	29.650000	1415000	29.6500
2015-12-14	29.600000	29.889999	28.850000	29.580000	2328600	29.5800
2015-12-15	29.680000	30.000000	26.459999	26.870001	5759200	26.8700
2015-12-16	26.889999	28.240000	26.260000	28.030001	2992100	28.0300
2015-12-17	28.139999	28.320000	27.190001	27.420000	1483900	27.4200
2015-12-18	27.309999	27.910000	26.900000	27.170000	1299800	27.1700
2015-12-21	27.170000	27.360001	26.030001	26.250000	1947600	26.2500
2015-12-22	26.250000	28.700001	26.150000	27.930000	2952700	27.9300
2015-12-23	27.950001	28.420000	27.440001	28.150000	1001000	28.1500
2015-12-24	28.270000	28.590000	27.900000	28.400000	587400	28.4000
2015-12-28	28.120001	28.379999	27.770000	27.879999	1004500	27.8799
2015-12-29	27.950001	28.540001	27.740000	28.480000	1103900	28.4800
2015-12-30	28.580000	28.780001	28.170000	28.250000	1068000	28.2500
2015-12-31	28.100000	28.969999	28.020000	28.799999	1301500	28.7999

date	profit	gain

```
2015-01-02  False   small_gain
2015-01-05  False  medium_gain
2015-01-06  False   small_gain
2015-01-07  False  medium_gain
2015-01-08   True     negative
2015-01-09   True     negative
2015-01-12  False  medium_gain
2015-01-13  False  medium_gain
2015-01-14  False   small_gain
2015-01-15  False  medium_gain
2015-01-16   True     negative
2015-01-20  False   small_gain
2015-01-21   True     negative
2015-01-22   True     negative
2015-01-23   True     negative
2015-01-26   True     negative
2015-01-27  False   small_gain
2015-01-28  False  medium_gain
2015-01-29   True     negative
2015-01-30  False   small_gain
2015-02-02   True     negative
2015-02-03   True     negative
2015-02-04   True     negative
2015-02-05  False   small_gain
2015-02-06  False  medium_gain
2015-02-09  False  medium_gain
2015-02-10   True     negative
2015-02-11   True     negative
2015-02-12   True     negative
2015-02-13  False   small_gain
...           ...          ...
2015-11-18   True     negative
2015-11-19  False   small_gain
2015-11-20   True     negative
2015-11-23  False   small_gain
2015-11-24   True     negative
2015-11-25   True     negative
2015-11-27  False   small_gain
2015-11-30   True     negative
2015-12-01   True     negative
2015-12-02   True     negative
2015-12-03  False   small_gain
2015-12-04  False   small_gain
2015-12-07  False   small_gain
2015-12-08   True     negative
2015-12-09  False   small_gain
2015-12-10   True     negative
2015-12-11  False  medium_gain
```

```
2015-12-14  False   small_gain
2015-12-15  False  medium_gain
2015-12-16   True     negative
2015-12-17  False   small_gain
2015-12-18  False   small_gain
2015-12-21  False   small_gain
2015-12-22   True     negative
2015-12-23   True     negative
2015-12-24   True     negative
2015-12-28  False   small_gain
2015-12-29   True     negative
2015-12-30  False   small_gain
2015-12-31   True     negative

[252 rows x 8 columns]
```

1.8 Grouping

An **extremely** powerful DataFrame method is **`groupby()`**.

This is entirely analagous to **`GROUP BY`** in SQL.

It will group the rows of a DataFrame by the values in one (or more) columns, and let you iterate through each group.

Here we will look at the average gain among the categories of gains (negative, small, medium and large) we defined above and stored in column `gain`.

```
In [307]: gain_groups = df.groupby('gain')
```

Essentially, **`gain_groups`** behaves like a dictionary * whose keys are the unique values found in the `gain` column, and * whose values are DataFrames that contain only the rows having the corresponding unique values.

```
In [308]: for gain, gain_data in gain_groups:
              print(gain)
              print(gain_data.head())
              print('============================')
```

```
medium_gain
                open       high        low      close   volume  adj_close  \
date
2015-01-05  54.540001  54.950001  52.330002  52.529999  2023000  52.529999
2015-01-07  53.320000  53.750000  51.759998  52.209999  1548200  52.209999
2015-01-12  56.000000  56.060001  53.430000  54.020000  2405100  54.020000
2015-01-13  54.470001  54.799999  52.520000  53.180000  1952100  53.180000
2015-01-15  53.000000  53.610001  50.029999  50.119999  2640400  50.119999

            profit         gain  test_column
date
2015-01-05  False  medium_gain  medium_gain
2015-01-07  False  medium_gain  medium_gain
```

```
2015-01-12  False  medium_gain  medium_gain
2015-01-13  False  medium_gain  medium_gain
2015-01-15  False  medium_gain  medium_gain
============================
negative
                open       high        low      close   volume  adj_close  \
date
2015-01-08  52.590000  54.139999  51.759998  53.830002  2015300  53.830002
2015-01-09  55.959999  56.990002  54.720001  56.070000  6222600  56.070000
2015-01-16  50.180000  51.490002  50.029999  51.389999  2183300  51.389999
2015-01-21  51.200001  53.500000  51.200001  53.410000  3248100  53.410000
2015-01-22  53.869999  55.279999  53.119999  54.799999  2295400  54.799999

           profit      gain test_column
date
2015-01-08   True  negative    negative
2015-01-09   True  negative    negative
2015-01-16   True  negative    negative
2015-01-21   True  negative    negative
2015-01-22   True  negative    negative
============================
small_gain
                open       high        low      close   volume  adj_close  \
date
2015-01-02  55.459999  55.599998  54.240002  55.150002  1664500  55.150002
2015-01-06  52.549999  53.930000  50.750000  52.439999  3762800  52.439999
2015-01-14  52.799999  53.680000  51.459999  52.200001  1854600  52.200001
2015-01-20  51.650002  51.779999  50.689999  51.410000  1227600  51.410000
2015-01-27  56.060001  56.160000  54.570000  55.630001  2410400  55.630001

           profit        gain test_column
date
2015-01-02  False  small_gain  small_gain
2015-01-06  False  small_gain  small_gain
2015-01-14  False  small_gain  small_gain
2015-01-20  False  small_gain  small_gain
2015-01-27  False  small_gain  small_gain
============================
```

```
In [309]: for gain, gain_data in df.groupby("gain"):
              print('The average closing value for the {} group is {}'.format(gain,
                                                                    gain_data.clos

The average closing value for the medium_gain group is 41.008888629629645
The average closing value for the negative group is 37.21357140476189
The average closing value for the small_gain group is 36.39636363636364
```

1.9 Other Pandas Classes

A DataFrame is essentially an annotated 2-D array.
Pandas also has annotated versions of 1-D and 3-D arrays.
A 1-D array in Pandas is called a **`Series`**.
A 3-D array in Pandas is called a **`Panel`**.
To use these, read the documentation!

1.10 Comparing multiple stocks

As a last task, we will use the experience we obtained so far – and learn some new things – in order to compare the performance of different stocks we obtained from Yahoo finance.

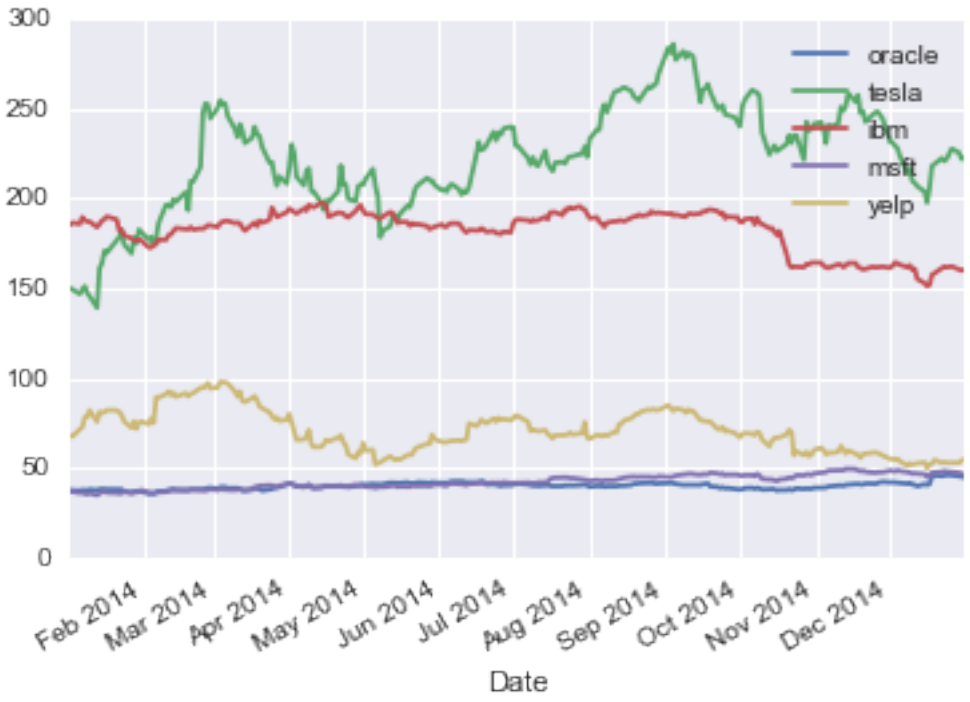
```
In [310]: stocks = ['ORCL', 'TSLA', 'IBM','YELP', 'MSFT']
          attr = 'Close'
          df = web.DataReader(stocks,
                              data_source,
                              start=datetime(2014, 1, 1),
                              end=datetime(2014, 12, 31))[attr]
          df.head()
```

Out[310]:

Date	IBM	MSFT	ORCL	TSLA	YELP
2014-01-02	185.529999	37.160000	37.840000	150.100006	67.919998
2014-01-03	186.639999	36.910000	37.619999	149.559998	67.660004
2014-01-06	186.000000	36.130001	37.470001	147.000000	71.720001
2014-01-07	189.710007	36.410000	37.849998	149.360001	72.660004
2014-01-08	187.970001	35.759998	37.720001	151.279999	78.419998

```
In [311]: df.ORCL.plot(label = 'oracle')
          df.TSLA.plot(label = 'tesla')
          df.IBM.plot(label = 'ibm')
          df.MSFT.plot(label = 'msft')
          df.YELP.plot(label = 'yelp')
          plt.legend(loc='best')
```

```
Out[311]: <matplotlib.legend.Legend at 0x11f2540f0>
```



Next, we will calculate returns over a period of length T, defined as:

$$r(t) = \frac{f(t) - f(t-T)}{f(t)}$$

The returns can be computed with a simple DataFrame method **`pct_change()`**. Note that for the first T timesteps, this value is not defined (of course):

```
In [312]: rets = df.pct_change(30)
          rets.iloc[25:35]
```

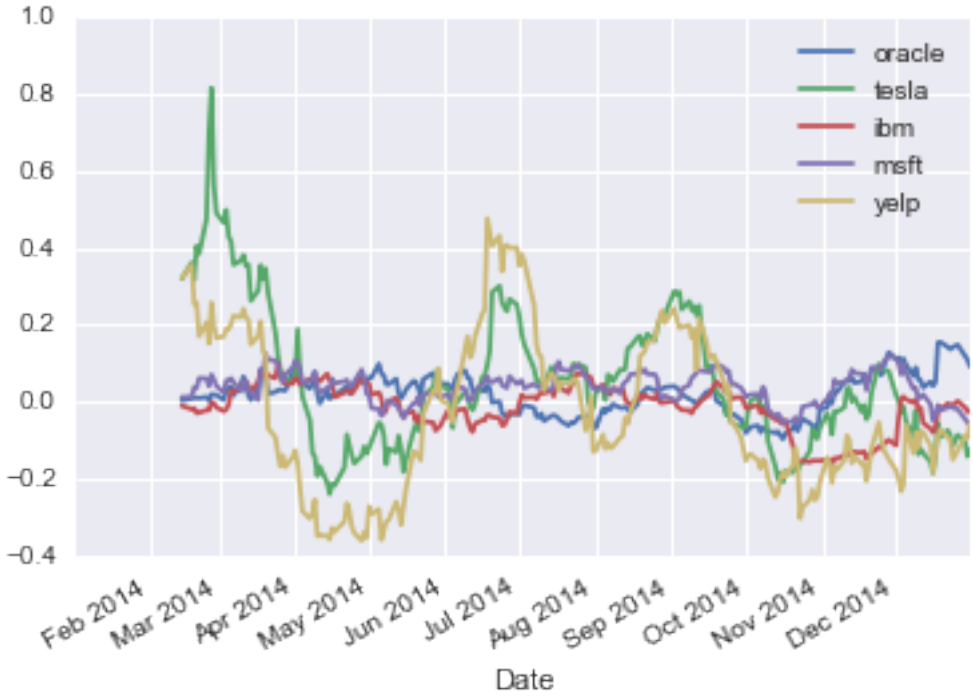
Out[312]:

Date	IBM	MSFT	ORCL	TSLA	YELP
2014-02-07	NaN	NaN	NaN	NaN	NaN
2014-02-10	NaN	NaN	NaN	NaN	NaN
2014-02-11	NaN	NaN	NaN	NaN	NaN
2014-02-12	NaN	NaN	NaN	NaN	NaN
2014-02-13	NaN	NaN	NaN	NaN	NaN
2014-02-14	-0.009918	0.012379	0.003700	0.320653	0.321849
2014-02-18	-0.018485	0.013817	0.009304	0.361995	0.355897
2014-02-19	-0.016398	0.038195	0.010675	0.317279	0.254880
2014-02-20	-0.028728	0.036803	0.011096	0.405798	0.257501
2014-02-21	-0.027558	0.062081	0.010074	0.385510	0.170875

Now we'll plot the timeseries of the returns of the different stocks.
Notice that the `NaN` values are gracefully dropped by the plotting function.

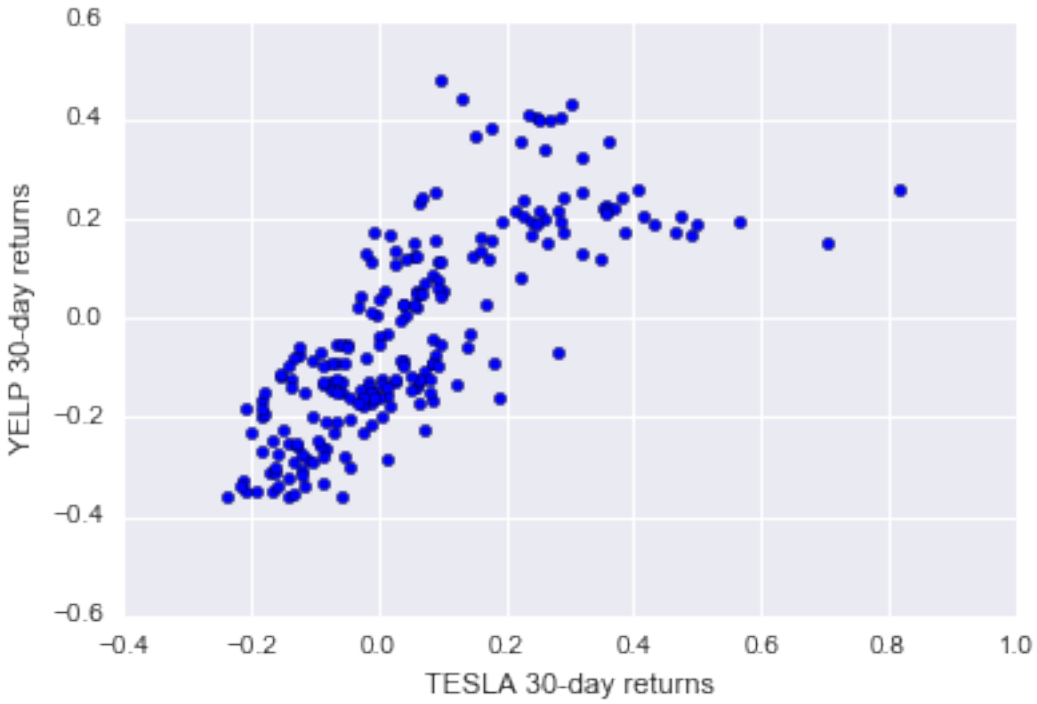
```
In [313]: rets.ORCL.plot(label = 'oracle')
          rets.TSLA.plot(label = 'tesla')
          rets.IBM.plot(label = 'ibm')
          rets.MSFT.plot(label = 'msft')
          rets.YELP.plot(label = 'yelp')
          plt.legend(loc='best')

Out[313]: <matplotlib.legend.Legend at 0x11f472be0>
```




```
In [314]: plt.scatter(rets.TSLA, rets.YELP)
          plt.xlabel('TESLA 30-day returns')
          plt.ylabel('YELP 30-day returns')

Out[314]: <matplotlib.text.Text at 0x11f4d7710>
```



There appears to be some (fairly strong) correlation between the movement of TSLA and YELP stocks. Let's measure this.

The correlation coefficient between variables X and Y is defined as follows:

$$\text{Corr}(X,Y) = \frac{E\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sigma_X \sigma_Y}$$

Pandas provides a DataFrame method to compute the correlation coefficient of all pairs of columns: **`corr()`**.

```
In [315]: rets.corr()
```

```
Out[315]:            IBM      MSFT      ORCL      TSLA      YELP
          IBM   1.000000  0.340422  0.025742  0.196504  0.104776
          MSFT  0.340422  1.000000  0.114090  0.415746  0.262595
          ORCL  0.025742  0.114090  1.000000  0.000981 -0.084176
          TSLA  0.196504  0.415746  0.000981  1.000000  0.768509
          YELP  0.104776  0.262595 -0.084176  0.768509  1.000000
```

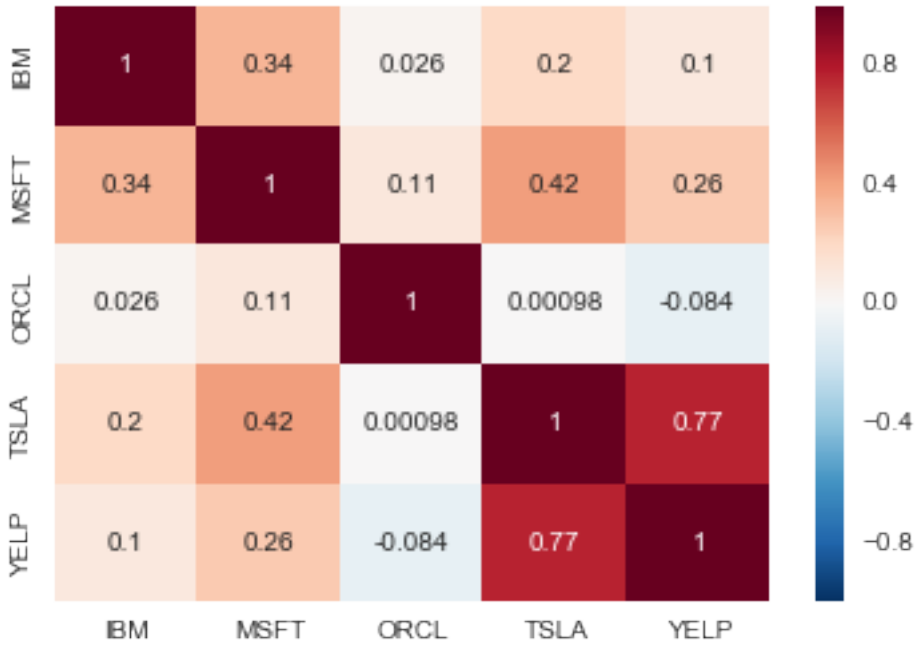
It takes a bit of time to examine that table and draw conclusions.

To speed that process up it helps to visualize the table.

We will learn more about visualization later, but for now this is a simple example.

```
In [316]: sns.heatmap(rets.corr(), annot=True)
```

```
Out[316]: <matplotlib.axes._subplots.AxesSubplot at 0x11f4def98>
```



Finally, it is important to know that the plotting performed by Pandas is just a layer on top of `matplotlib` (i.e., the `plt` package).

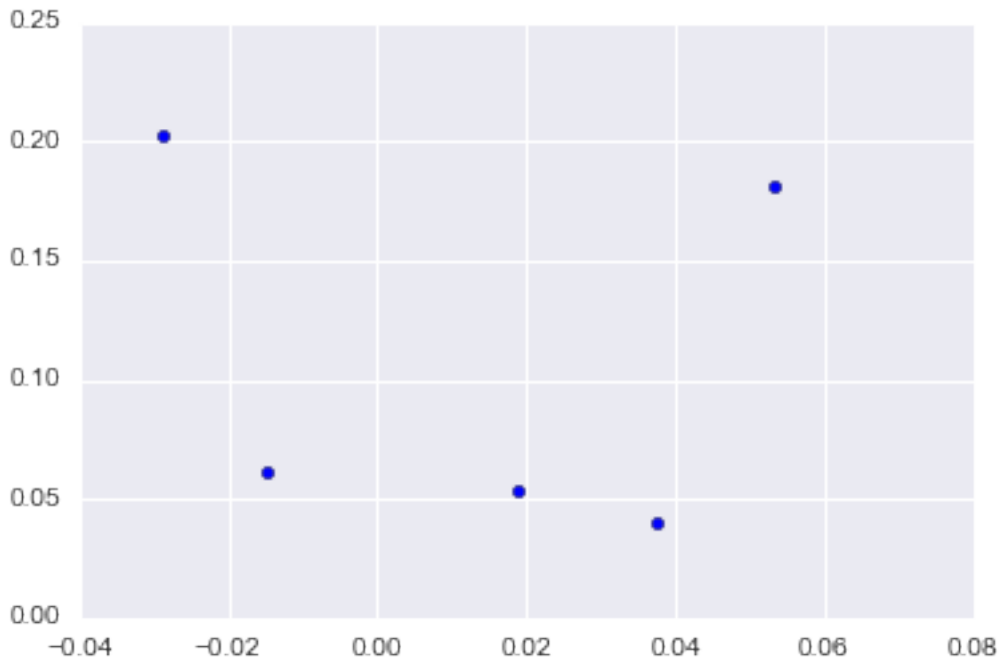
So Panda's plots can (and should) be replaced or improved by using additional functions from `matplotlib`.

For example, suppose we want to know both the returns as well as the standard deviation of the returns of a stock (i.e., its risk).

Here is visualization of the result of such an analysis, and we construct the plot using only functions from `matplotlib`.

```
In [317]: plt.scatter(rets.mean(), rets.std())
          # plt.xlabel('Expected returns')
          # plt.ylabel('Standard Deviation (Risk)')
          # for label, x, y in zip(rets.columns, rets.mean(), rets.std()):
          #    plt.annotate(
          #        label,
          #        xy = (x, y), xytext = (20, -20),
          #        textcoords = 'offset points', ha = 'right', va = 'bottom',
          #        bbox = dict(boxstyle = 'round,pad=0.5', fc = 'yellow', alpha =
          #        arrowprops = dict(arrowstyle = '->', connectionstyle = 'arc3,rad

Out[317]: <matplotlib.collections.PathCollection at 0x11e650e48>
```



To understand what these functions are doing, (especially the `annotate` function), you will need to consult the online documentation for matplotlib. Just use Google to find it.