2 描述性统计及其Python应用

1. 读取csv格式的数据

首先，加载数据文件到内存里

```
import pandas
pandas.read_csv("datasets/Facebook.csv", index_col=0)
```

1. 读取csv格式的数据

Date	High	Low	Open	Close	Volume	Adj Close
2017-01-03	12.60	12.13	12.20	12.59	40510800.0	10.255996
2017-01-04	13.27	12.74	12.77	13.17	77638100.0	10.728471
2017-01-05	13.22	12.63	13.21	12.77	75628400.0	10.402627
2017-01-06	12.84	12.64	12.80	12.76	40315900.0	10.394479
2017-01-09	12.86	12.63	12.79	12.63	39438400.0	10.288579
...	...	...	...	...	...	...
2021-09-20	13.19	12.63	13.07	12.82	126152200.0	12.820000
2021-09-21	12.97	12.66	12.89	12.77	60473100.0	12.770000
2021-09-22	13.39	12.91	12.92	13.23	75784000.0	13.230000
2021-09-23	13.74	13.32	13.35	13.71	68708100.0	13.710000
2021-09-24	13.94	13.64	13.66	13.78	57440200.0	13.780000

1191 rows × 6 columns

它返回的是一个pandas.DataFrame的对象，称之为“数据框”类型。

2. 基础描述性统计信息

```
df.describe()
```

	High	Low	Open	Close	Volume	Adj Close
count	1191.000000	1191.000000	1191.000000	1191.000000	1.191000e+03	1191.000000
mean	10.225558	9.985802	10.113468	10.102712	5.308135e+07	9.394648
std	2.269341	2.251395	2.261142	2.262616	2.983678e+07	2.043040
min	4.420000	3.960000	4.270000	4.010000	9.549600e+06	4.010000
25%	8.895000	8.715000	8.810000	8.800000	3.319940e+07	8.369491
50%	10.260000	10.070000	10.170000	10.190000	4.456190e+07	9.287052
75%	11.830000	11.620000	11.715000	11.715000	6.391495e+07	10.318997
max	16.450001	15.800000	16.330000	15.990000	2.823941e+08	15.990000

2. 基础描述性统计信息

count：个数

mean：均值

std：标准差

min：最小值

25%：分位数为25%的数值

50%：分位数为50%的数值

75%：分位数为75%的数值

max：最大值

```
df.describe()
```

	High	Low	Open	Close	Volume	Adj Close
count	1191.000000	1191.000000	1191.000000	1191.000000	1.191000e+03	1191.000000
mean	10.225558	9.985802	10.113468	10.102712	5.308135e+07	9.394648
std	2.269341	2.251395	2.261142	2.262616	2.983678e+07	2.043040
min	4.420000	3.960000	4.270000	4.010000	9.549600e+06	4.010000
25%	8.895000	8.715000	8.810000	8.800000	3.319940e+07	8.369491
50%	10.260000	10.070000	10.170000	10.190000	4.456190e+07	9.287052
75%	11.830000	11.620000	11.715000	11.715000	6.391495e+07	10.318997
max	16.450001	15.800000	16.330000	15.990000	2.823941e+08	15.990000

分位数

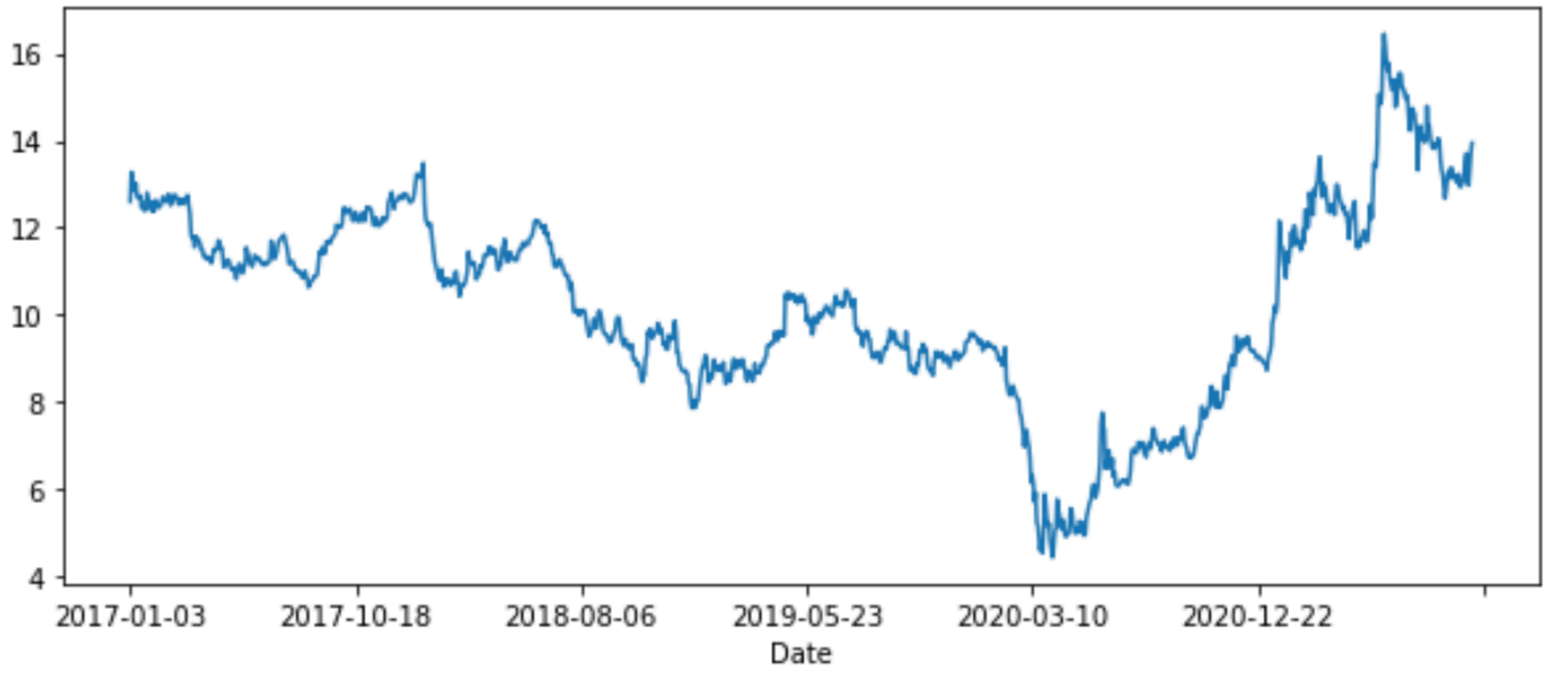
一组数据：[5, 2, 4 ..., 1] 通过从小到大排序后， [1, 2, 4, ... 5]

折线图

我们可以通过绘制折线图来了解数据框中的单个属性来观察其数据趋势：

```
df['High'].plot(figsize=(10, 4))
```

```
<AxesSubplot:xlabel='Date'>
```



频数图

或者通过绘制频数分布图（条状图）来了解数据的一个分布情况：

```
df['High'].hist(bins=50, figsize=(10, 4))
```

```
<AxesSubplot:>
```