FINAL PROJECT

IDS
UE18CS203

TEAM

- PES1201801886 ,Gaganmanohar T
- PES1201802120 ,Sandeep Kumar S
- PES1201800384 ,M S Sai Srikar

TOPICS COVERED

- INTRODUCTION
- DATA CLEANING
- NORMALIZATION
- CORRELATIONS
- HYPOTHESIS TESTING

INTRODUCTION

- The dataset is basically the stock parameters of Google.inc over 13 years span.

Index	Date	Open	High	Low	Close	Volume	Ex-Dividend
0	2004-08-19	100.01	104.06	95.96	100.335	4.4659e+07	0
1	2004-08-20	101.01	109.08	100.5	108.31	2.28343e+07	0
2	2004-08-23	110.76	113.48	109.05	109.4	1.82561e+07	0
3	2004-08-24	111.24	111.6	103.57	104.87	1.52473e+07	0
4	2004-08-25	104.76	nan	103.88	106	9.1886e+06	0
5	2004-08-26	104.95	107.95	104.66	107.91	7.0948e+06	0
6	2004-08-27	108.1	108.62	105.69	106.15	6.2117e+06	0
7	2004-08-30	105.28	105.49	nan	102.01	5.1967e+06	0
8	2004-08-31	102.32	103.71	102.16	102.37	4.9178e+06	0
9	2004-09-01	102.7	102.97	99.67	100.25	nan	0
10	2004-09-02	99.09	102.37	98.94	101.51	1.51186e+07	0
11	2004-09-03	100.95	101.74	99.32	100.01	5.1524e+06	0
12	2004-09-07	nan	102	99.61	101.58	5.8475e+06	0
13	2004-09-08	100.74	nan	100.5	102.3	4.9856e+06	0
14	2004-09-09	102.5	102.71	nan	102.31	4.0617e+06	0
15	2004-09-10	101.47	106.56	101.3	105.33	8.6988e+06	0
16	2004-09-13	nan	108.41	106.46	107.5	7.8441e+06	0
17	2004-09-14	107.44	112	106.79	111.49	1.08289e+07	0
18	2004-09-15	110.56	114.23	110.2	112	1.0713e+07	0
19	2004-09-16	112.34	115.8	111.65	113.97	9.2663e+06	0
20	2004-09-17	114.42	117.49	113.55	117.49	9.4725e+06	0
21	2004-09-20	116.95	121.6	116.77	119.36	1.06287e+07	0
22	2004-09-21	120.2	120.42	117.51	117.84	7.2287e+06	0
23	2004-09-22	117.45	119.67	nan	118.38	7.5812e+06	0
24	2004-09-23	118.84	122.63	117.02	120.82	8.5356e+06	0

ATTRIBUTES

- Definitions
- Date: It defines the time where the respective values are defined.
- Open: The price when the market opened in the morning.
- High: The highest price during that trading day
- Low: The lowest price during that trading day.
- Close: The price when the market closed in the afternoon
- Volume: It is commonly reported as the number of shares that changed hands during a given day.
- Adj Values: A price adjusted to make prices comparable over time

Open	High	Low	Close	Volume
100.01	104.06	95.96	100.335	4.4659e+07

DATA CLEANING

- Data cleaning contains the following process
  A.Identifying the Null values in the data set.
  B.Filling the values in place of Null value cells

- In our dataset we replaced the null values with the average value of the respective attribute.

NULL VALUES

Index	Date	Open	High	Low	Close	Volume	Ex-Dividend
58	False	False	False	False	False	False	False
59	False	False	False	False	False	False	False
60	False	False	True	False	False	False	False
61	False	False	False	False	False	False	False
62	False	False	False	False	False	False	False
63	False	False	False	False	False	False	False
64	False	False	False	False	False	False	False
65	False	False	False	False	False	True	False
66	False	False	False	False	False	False	False
67	False	False	False	False	False	False	False
68	False	False	False	False	False	False	False
69	False	False	False	False	False	False	False
70	False	False	False	False	False	False	False
71	False	False	False	False	False	False	False
72	False	False	False	False	False	False	False
73	False	False	False	False	False	False	False
74	False	False	False	False	False	False	False
75	False	False	True	True	False	False	False
76	False	False	False	False	False	False	False
77	False	False	False	False	True	False	False
78	False	False	False	False	False	False	False

Number of null values

```
In [68]: data.isnull().sum()
Out[68]:
Date            0
Open           28
High           60
Low            59
Close          55
Volume         52
Ex-Dividend     0
Split Ratio     0
Adj. Open       0
Adj. High       0
Adj. Low        0
Adj. Close      0
Adj. Volume     0
dtype: int64

In [69]:
```

Replacing null values

Index	Open
0	100.01
1	101.01
2	110.76
3	111.24
4	104.76
5	104.95
6	108.1
7	105.28
8	102.32
9	102.7
10	99.09
11	100.95
12	561.49
13	100.74
14	102.5
15	101.47
16	561.49
17	107.44
18	110.56
19	112.34
20	114.42
21	116.95
22	120.2
23	117.45
24	118.84
25	120.97

Index	Close
106	193.92
107	188.28
108	180.72
109	177.12
110	189.24
111	188.08
112	190.34
113	195.62
114	191.9
115	561.107
116	210.86
117	204.36
118	196.03
119	198.64
120	191.58
121	187.98
122	187.4
123	192.99
124	195.23
125	198.41
126	197.9
127	197.95
128	191.37
129	193.95
130	188.89

Index	High
0	104.06
1	109.08
2	113.48
3	111.6
4	566.898
5	107.95
6	108.62
7	105.49
8	103.71
9	102.97
10	102.37
11	101.74
12	102
13	566.898
14	102.71
15	106.56
16	108.41
17	112
18	114.23
19	115.8
20	117.49
21	121.6
22	120.42
23	119.67
24	122.63
25	124.1

Index	Low
0	95.96
1	100.5
2	109.05
3	103.57
4	103.88
5	104.66
6	105.69
7	555.129
8	102.16
9	99.67
10	98.94
11	99.32
12	99.61
13	100.5
14	555.129
15	101.3
16	106.46
17	106.79
18	110.2
19	111.65
20	113.55
21	116.77
22	117.51
23	555.129
24	117.02
25	119.76

Index	Close
21	119.36
22	117.84
23	118.38
24	120.82
25	561.107
26	118.26
27	126.86
28	131.08
29	129.6
30	132.58

NORMALIZATION

Before

Index	Open	High	Low	Close	Volume	Ex-Dividend	Split Ratio	Adj. Open	Adj. High	Adj. Low	Adj. Close	Adj. Volume
0	100.01	104.06	95.96	100.335	4.4659e+07	0	1	50.1598	52.1911	48.1286	50.3228	44659000
1	101.01	109.08	100.5	108.31	2.28343e+07	0	1	50.6614	54.7089	50.4056	54.3227	22834300
2	110.76	113.48	109.05	109.4	1.82561e+07	0	1	55.5515	56.9157	54.6938	54.8694	18256100
3	111.24	111.6	103.57	104.87	1.52473e+07	0	1	55.7922	55.9728	51.9454	52.5974	15247300
4	104.76	566.898	103.88	106	9.1886e+06	0	1	52.5422	54.1672	52.1008	53.1641	9188600
5	104.95	107.95	104.66	107.91	7.0948e+06	0	1	52.6375	54.1421	52.492	54.1221	7094800
6	108.1	108.62	105.69	106.15	6.2117e+06	0	1	54.2174	54.4782	53.0086	53.2393	6211700
7	105.28	105.49	555.129	102.01	5.1967e+06	0	1	52.803	52.9083	51.1629	51.1629	5196700
8	102.32	103.71	102.16	102.37	4.9178e+06	0	1	51.3184	52.0156	51.2382	51.3435	4917800
9	102.7	102.97	99.67	100.25	nan	0	1	51.509	51.6444	49.9893	50.2802	9138200
10	99.09	102.37	98.94	101.51	1.51186e+07	0	1	49.6984	51.3435	49.6232	50.9122	15118600
11	100.95	101.74	99.32	100.01	5.1524e+06	0	1	50.6313	51.0275	49.8138	50.1598	5152400
12	561.49	102	99.61	101.58	5.8475e+06	0	1	50.6614	51.1579	49.9592	50.9473	5847500
13	100.74	566.898	100.5	102.3	4.9856e+06	0	1	50.526	51.6745	50.4056	51.3084	4985600
14	102.5	102.71	555.129	102.31	4.0617e+06	0	1	51.4087	51.514	50.6564	51.3134	4061700
15	101.47	106.56	101.3	105.33	8.6988e+06	0	1	50.8921	53.445	50.8068	52.8281	8698800
16	561.49	108.41	106.46	107.5	7.8441e+06	0	1	53.4801	54.3728	53.3948	53.9164	7844100
17	107.44	112	106.79	111.49	1.08289e+07	0	1	53.8863	56.1734	53.5603	55.9176	10828900
18	110.56	114.23	110.2	112	1.0713e+07	0	1	55.4512	57.2919	55.2706	56.1734	10713000
19	112.34	115.8	111.65	113.97	9.2663e+06	0	1	56.3439	58.0793	55.9979	57.1615	9266300
20	114.42	117.49	113.55	117.49	9.4725e+06	0	1	57.3871	58.9269	56.9508	58.9269	9472500
21	116.95	121.6	116.77	119.36	1.06287e+07	0	1	58.6561	60.9883	58.5658	59.8648	10628700
22	120.2	120.42	117.51	117.84	7.2287e+06	0	1	60.2861	60.3964	58.9369	59.1024	7228700
23	117.45	119.67	555.129	118.38	7.5812e+06	0	1	58.9068	60.0203	58.5858	59.3733	7581200
24	118.84	122.63	117.02	120.82	8.5356e+06	0	1	59.604	61.5049	58.6912	60.5971	8535600
25	120.97	124.1	119.76	561.107	9.1234e+06	0	1	60.6723	62.2421	60.0654	60.1005	9123400

After normalisation

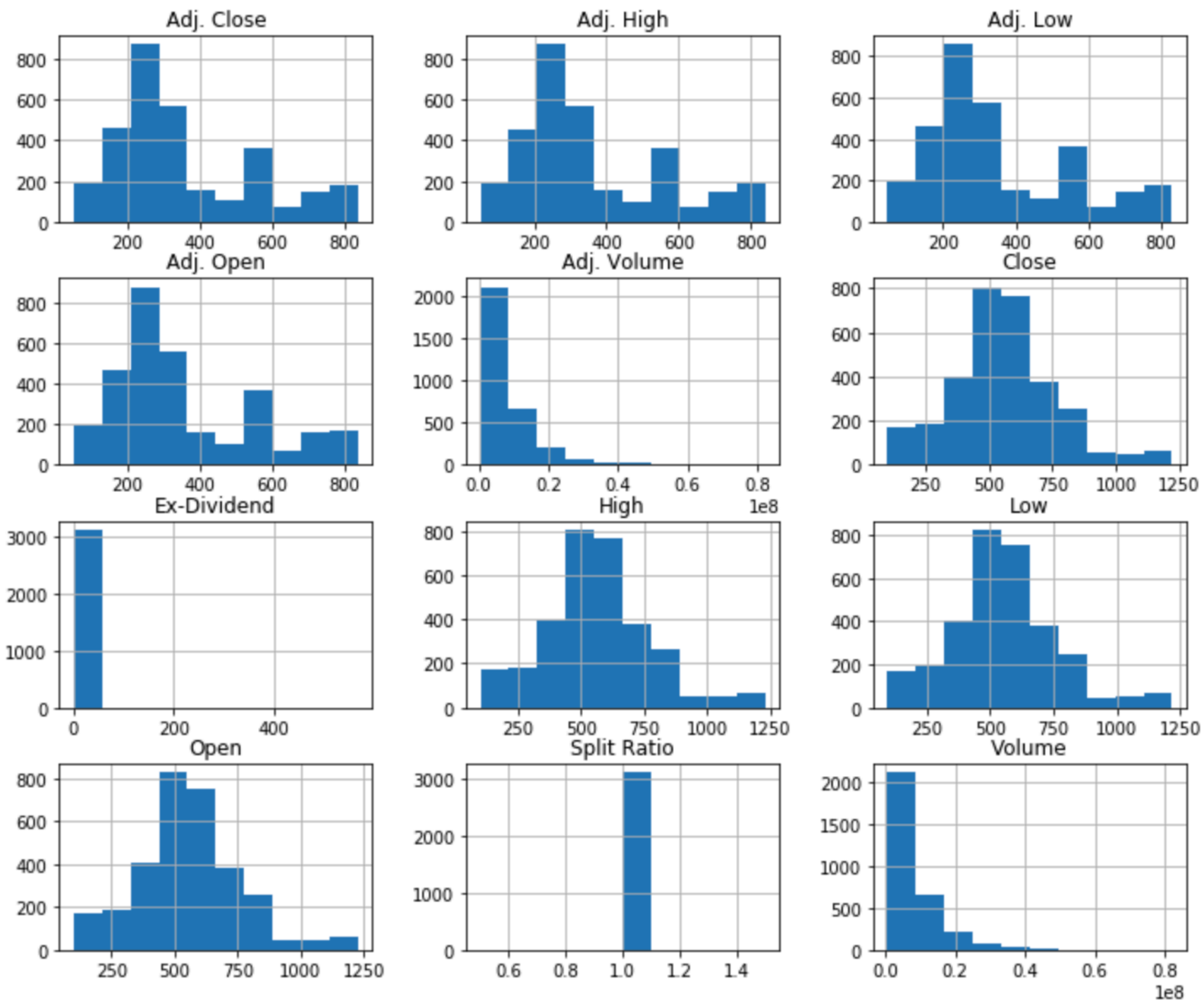
Index	Open	High	Low	Close	Volume	Ex-Dividend	Split Ratio	Adj. Open	Adj. High	Adj. Low	Adj. Close	Adj. Volume
0	-2.30651	-2.32759	-2.32404	-2.31211	4.30344	-0.0178885	nan	-1.56055	-1.5579	-1.56384	-1.55971	4.31615
1	-2.30151	-2.30234	-2.30106	-2.27209	1.71275	-0.0178885	nan	-1.55798	-1.54508	-1.55209	-1.53921	1.71788
2	-2.25278	-2.28022	-2.25778	-2.26662	1.1693	-0.0178885	nan	-1.53294	-1.53384	-1.52996	-1.53641	1.17283
3	-2.25038	-2.28967	-2.28552	-2.28935	0.81214	-0.0178885	nan	-1.5317	-1.53864	-1.54414	-1.54805	0.814628
4	-2.28277	-1.71518e-15	-2.28395	-2.28368	0.0929456	-0.0178885	nan	-1.54835	-1.54784	-1.54334	-1.54515	0.0933261
5	-2.28182	-2.30803	-2.28	-2.2741	-0.155598	-0.0178885	nan	-1.54786	-1.54797	-1.54132	-1.54024	-0.155945
6	-2.26607	-2.30466	-2.27479	-2.28293	-0.260426	-0.0178885	nan	-1.53977	-1.54625	-1.53866	-1.54476	-0.26108
7	-2.28017	-2.3204	1.15083e-15	-2.3037	-0.380911	-0.0178885	nan	-1.54701	-1.55425	-1.54818	-1.5554	-0.381918
8	-2.29496	-2.32935	-2.29266	-2.3019	-0.414017	-0.0178885	nan	-1.55462	-1.55879	-1.54779	-1.55448	-0.415122
9	-2.29306	-2.33307	-2.30526	-2.31253	nan	-0.0178885	nan	-1.55364	-1.56068	-1.55424	-1.55993	0.0873259
10	-2.31111	-2.33609	-2.30896	-2.30621	0.796863	-0.0178885	nan	-1.56291	-1.56222	-1.55613	-1.55669	0.799306
11	-2.30181	-2.33926	-2.30703	-2.31374	-0.386169	-0.0178885	nan	-1.55814	-1.56383	-1.55514	-1.56055	-0.387192
12	-1.13643e-15	-2.33795	-2.30556	-2.30586	-0.303658	-0.0178885	nan	-1.55798	-1.56316	-1.55439	-1.55651	-0.304439
13	-2.30286	-1.71518e-15	-2.30106	-2.30225	-0.405969	-0.0178885	nan	-1.55868	-1.56053	-1.55209	-1.55466	-0.40705
14	-2.29406	-2.33438	1.15083e-15	-2.3022	-0.51564	-0.0178885	nan	-1.55415	-1.56135	-1.55079	-1.55463	-0.517042
15	-2.29921	-2.31502	-2.29701	-2.28704	0.0348041	-0.0178885	nan	-1.5568	-1.55152	-1.55002	-1.54687	0.0350143
16	-1.13643e-15	-2.30571	-2.27089	-2.27615	-0.0666525	-0.0178885	nan	-1.54355	-1.54679	-1.53667	-1.54129	-0.0667395
17	-2.26937	-2.28766	-2.26922	-2.25613	0.287656	-0.0178885	nan	-1.54147	-1.53762	-1.53581	-1.53104	0.288607
18	-2.25378	-2.27644	-2.25196	-2.25357	0.273899	-0.0178885	nan	-1.53345	-1.53193	-1.52699	-1.52973	0.274809
19	-2.24488	-2.26855	-2.24463	-2.24369	0.102169	-0.0178885	nan	-1.52888	-1.52792	-1.52324	-1.52467	0.102576
20	-2.23449	-2.26005	-2.23501	-2.22602	0.126646	-0.0178885	nan	-1.52354	-1.5236	-1.51832	-1.51562	0.127125
21	-2.22184	-2.23938	-2.21871	-2.21664	0.263892	-0.0178885	nan	-1.51704	-1.5131	-1.50999	-1.51081	0.264773
22	-2.2056	-2.24531	-2.21497	-2.22427	-0.139703	-0.0178885	nan	-1.50869	-1.51612	-1.50807	-1.51472	-0.140004
23	-2.21934	-2.24909	1.15083e-15	-2.22156	-0.0978599	-0.0178885	nan	-1.51575	-1.51803	-1.50988	-1.51333	-0.0980384
24	-2.2124	-2.2342	-2.21745	-2.20932	0.0154316	-0.0178885	nan	-1.51218	-1.51047	-1.50934	-1.50706	0.015585
25	-2.20175	-2.22681	-2.20358	1.7114e-15	0.085206	-0.0178885	nan	-1.50671	-1.50672	-1.50225	-1.50961	0.0855639

GRAPH VISUALISATION

Different kinds of plots are used over our attributes in order to analyse the dataset. Few kinds are line plot, bar graphs and co relation plots.

BAR GRAPHS

```
      <matplotlib.axes._subplots.AxesSubplot object at 0x1a1773a4a8>]],
dtype=object)
```



LINE GRAPHS

- We used line graphs to analyse the trends.
- This is because the dataset contains the data of long period of time(2004- 2017).
- In order to analyse trends over long period of time line graphs are the best.

Date vs Stock High Price

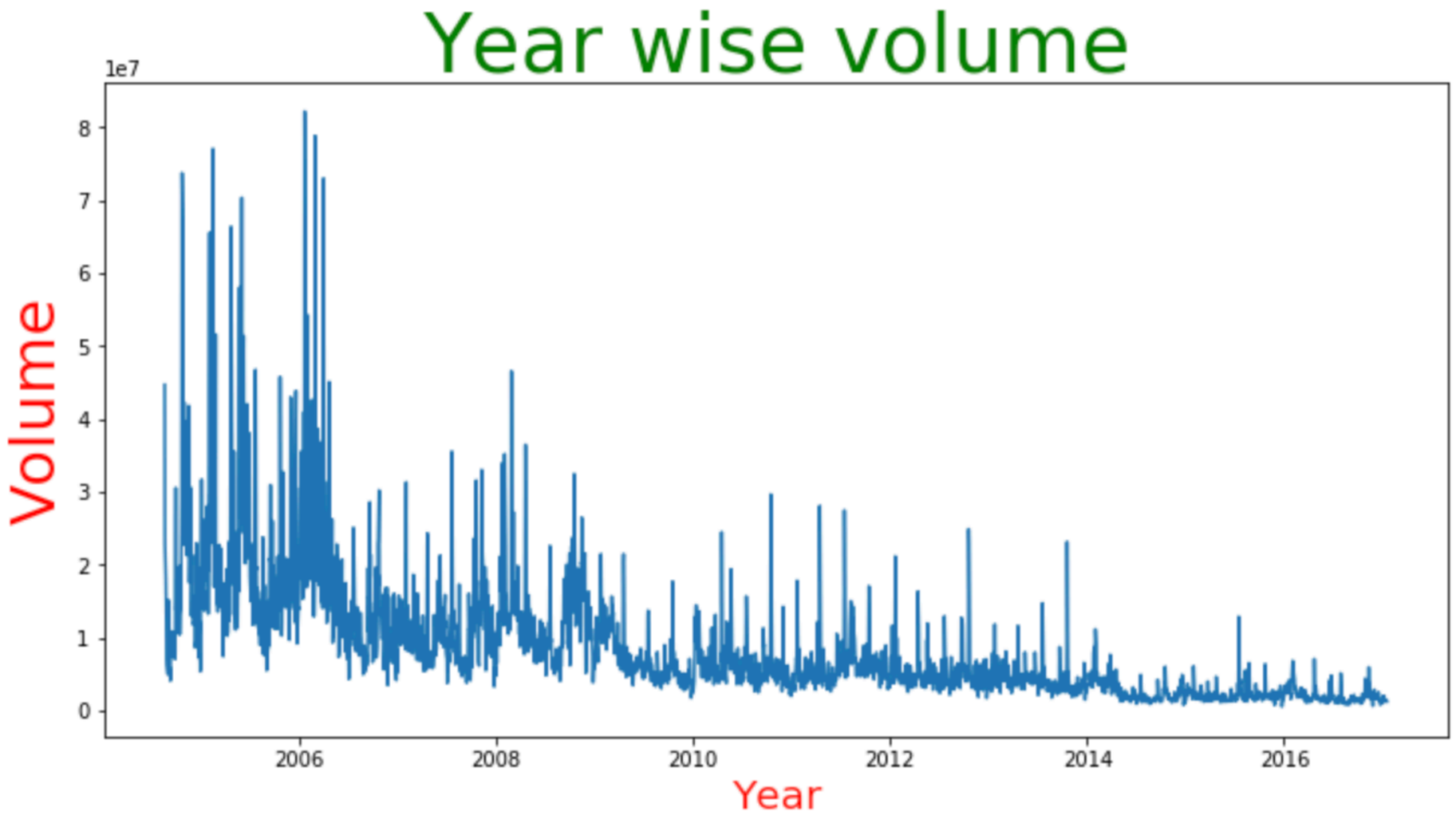
```
plt.plot(data['High'])
plt.title("Date vs Stock High Price",fontsize=40,color='g')
plt.xlabel("Date",fontsize=20,color='r')
plt.ylabel("Stock High Price",fontsize=30,color='r')
plt.show()
```

- Regarding stock High price, the graph increases till 2008 and it faces a downfall. This was due to the economic recession which occurred during that period.

- The major downfall was experienced in 2014.

Year Wise Volume

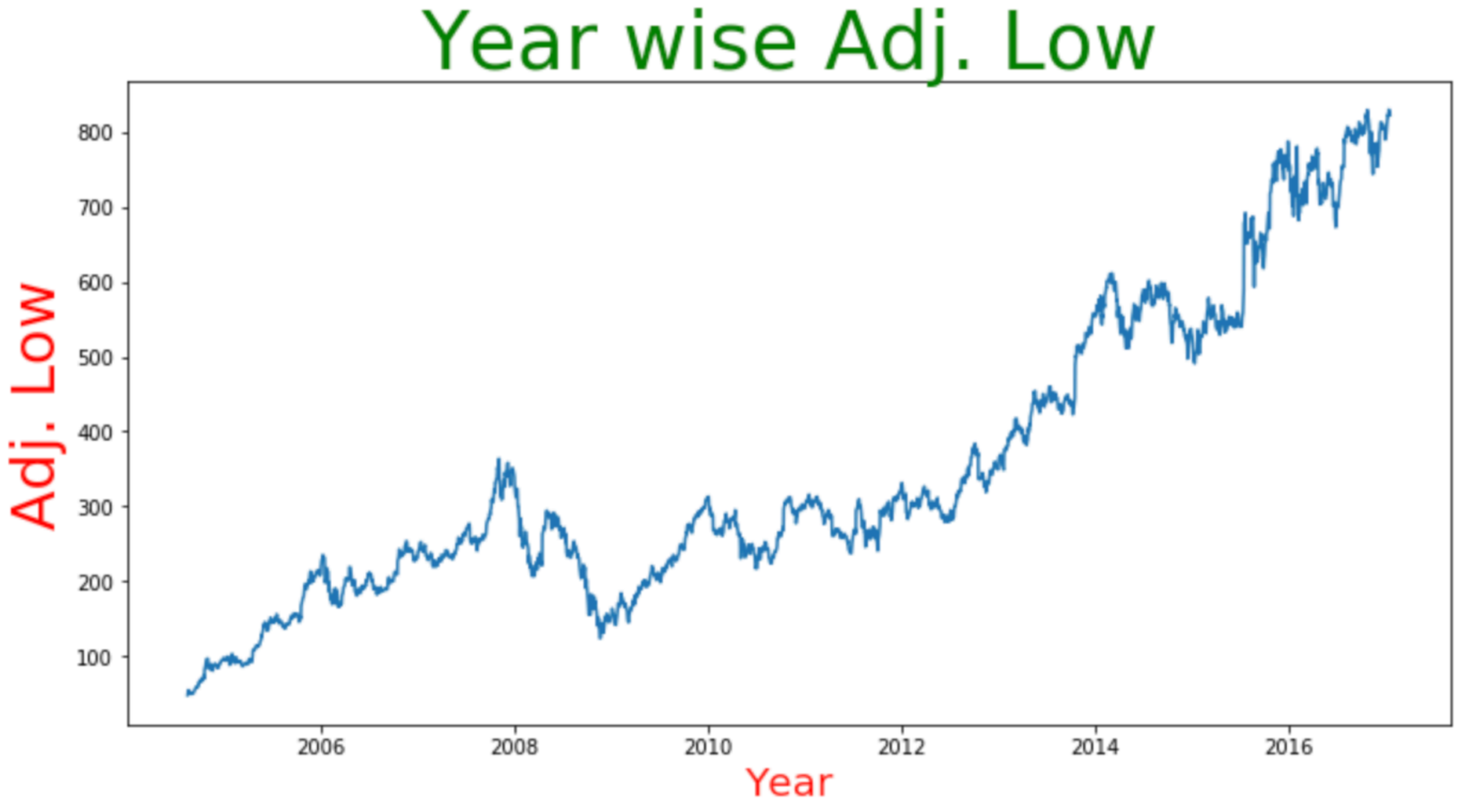
In [11]:
```
plt.figure(figsize=(12,6))
#Volume Ex-Dividend Split Ratio Adj. Open Adj. High Adj. Low Adj. Close Adj. Volume
plt.plot(data['Volume'])
plt.title("Year wise volume",fontsize=40,color='g')
plt.xlabel("Year",fontsize=20,color='r')
plt.ylabel("Volume",fontsize=30,color='r')
plt.show()
```



- The volume has been continuously decreasing over years.
- This shows that shareholders are holding on to their stock instead of selling them.
- This shows that the company is performing well over years.

Year wise Adj.low

```
plt.title("Year wise Adj. Low",fontsize=40,color='g')
plt.xlabel("Year",fontsize=20,color='r')
plt.ylabel("Adj. Low",fontsize=30,color='r')
plt.show()
```

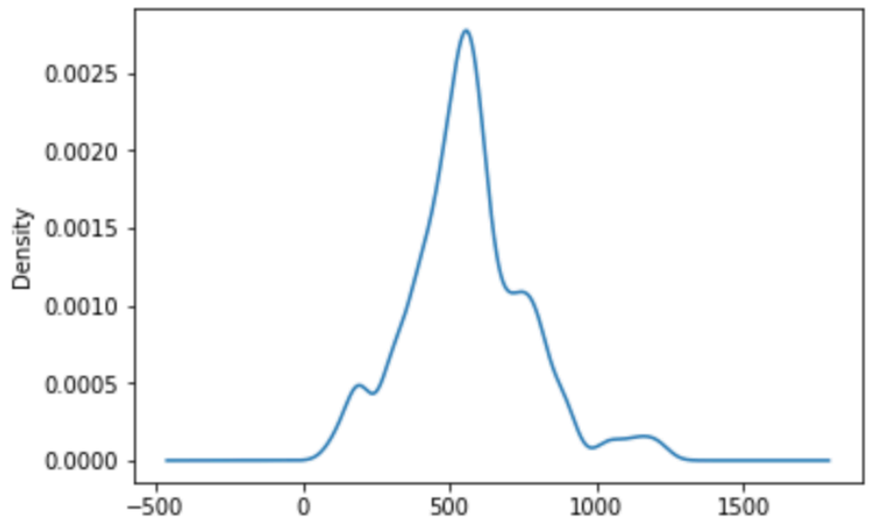


- The stock Low is increasing, this shows that the value of the share is increasing over years.

kde Plots

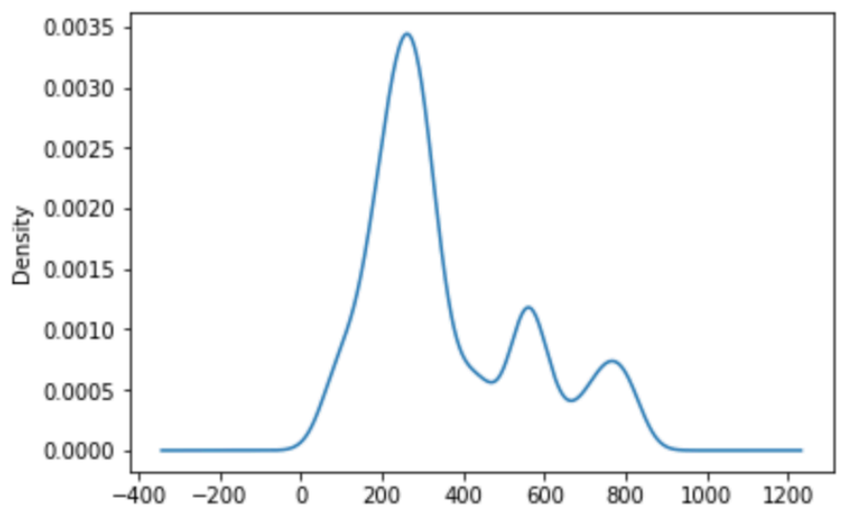
```
In [13]: data['High'].plot("kde")
```

```
Out[13]: <matplotlib.axes._subplots.AxesSubplot at 0x1037214e0>
```

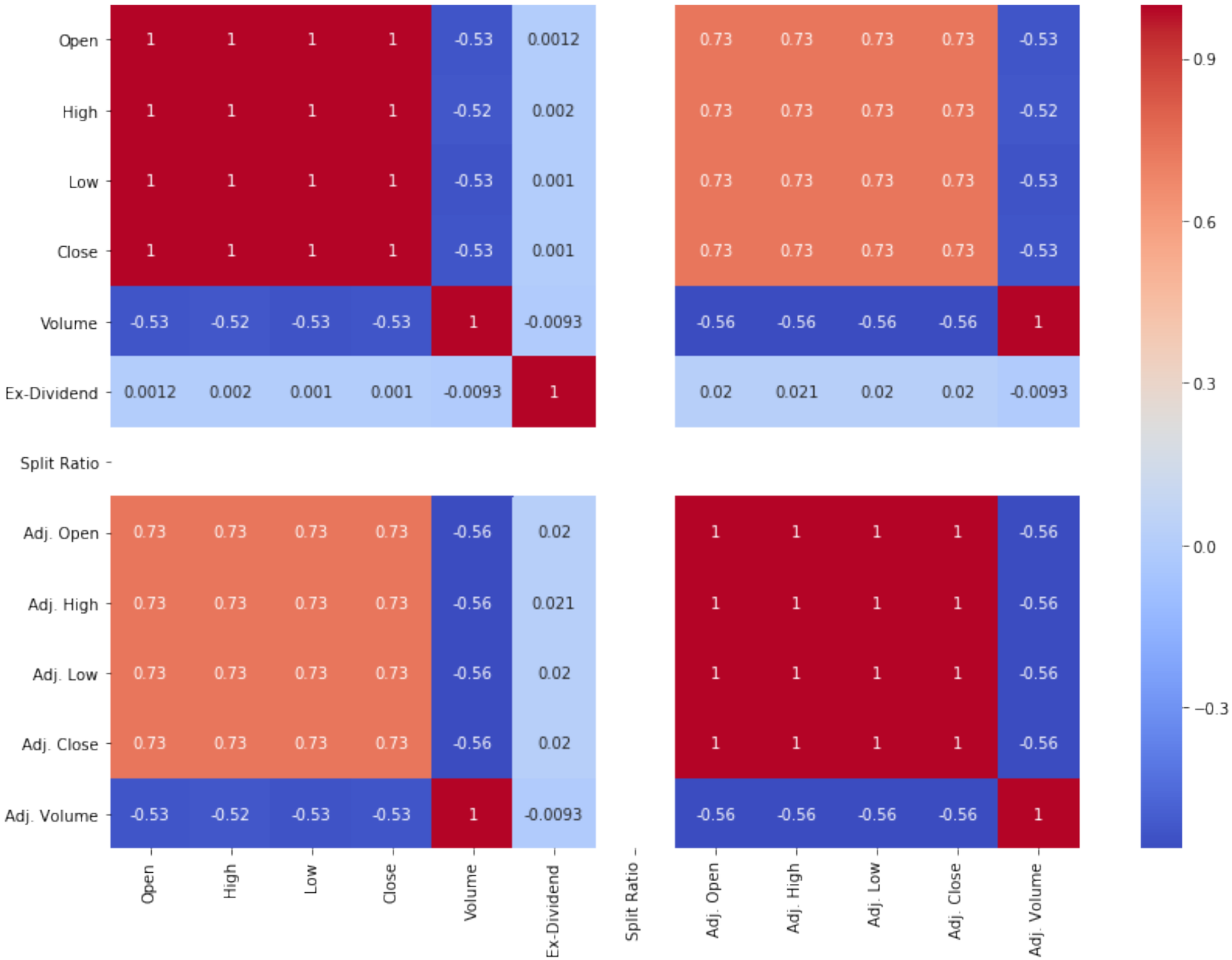



```
In [14]: data['Adj. High'].plot("kde")
```

```
Out[14]: <matplotlib.axes._subplots.AxesSubplot at 0x103690a58>
```



Correlations



Hypothesis Testing

On "Open" Attribute

In [30]:
```
from scipy.stats import norm
from math import sqrt
a=(ds[1:31]['Open'])
def one_sided_hypo(sample_mean,pop_mean,std_dev,sample_size,alpha):
    actual_z = abs(norm.ppf(alpha))
    hypo_z = (sample_mean - pop_mean) / (std_dev/sqrt(sample_size))
    print('actual z value : ', hypo_z, '\n')
    if hypo_z >= actual_z:
        return True
    else:
        return False

alpha = 0.05
sample_size= 30
sample_mean = a.mean()
pop_mean = ds['Open'].mean()
std_dev = a.std()

print("H0: Average duration of talks is less than or equal to population mean")
print("H1: Average duration of talks is grester than population mean")
print('H0: μ <=', pop_mean)
print('H1: μ >', pop_mean)
print('alpha value is: ', alpha, '\n')
reject = one_sided_hypo(sample_mean, pop_mean, std_dev, sample_size, alpha)
if reject:
    print('Reject NULL hypothesis')
else:
    print('Failed to reject NULL hypothesis')
```

```
H0: Average duration of talks is less than or equal to population mean
H1: Average duration of talks is grester than population mean
H0: μ <= 560.1715665600002
H1: μ > 560.1715665600002
alpha value is:  0.05

actual z value :  -264.6198746687786

Failed to reject NULL hypothesis
```

Hypothesis Testing

```
a=(ds[1:31]['Close'])
def one_sided_hypo(sample_mean,pop_mean,std_dev,sample_size,alpha):
    actual_z = abs(norm.ppf(alpha))
    hypo_z = (sample_mean - pop_mean) / (std_dev/sqrt(sample_size))
    print('actual z value : ', hypo_z, '\n')
    if hypo_z >= actual_z:
        return True
    else:
        return False

alpha = 0.05
sample_size= 30
sample_mean = a.mean()
pop_mean = ds['Close'].mean()
std_dev = a.std()

print("H0: Average duration of talks is less than or equal to population mean")
print("H1: Average duration of talks is grester than population mean")
print('H0: μ <=', pop_mean)
print('H1: μ >', pop_mean)
print('alpha value is: ', alpha, '\n')
reject = one_sided_hypo(sample_mean, pop_mean, std_dev, sample_size, alpha)
if reject:
    print('Reject NULL hypothesis')
else:
    print('Failed to reject NULL hypothesis')
```

```
H0: Average duration of talks is less than or equal to population mean
H1: Average duration of talks is grester than population mean
H0: μ <= 559.9073419199997
H1: μ > 559.9073419199997
alpha value is:  0.05

actual z value :  -251.37099114493964

Failed to reject NULL hypothesis
```

On “Close” Attribute

Conclusion

- From correlation map, we can identify the relation between different attributes.

- It is observed that the normal values(eg. High,Low,Open and Close) weighs almost twice the adjusted values(eg. adj.High, adj.Low, adj.Open)