

## Data Analytics Project

### Stock-taking of data

Bitcoin\_price.csv

Number of columns: 7

Number of rows: 1760

Attributes:

```
In [2]: df.dtypes

Out[2]: Date          datetime64[ns]
       Open           float64
       High           float64
       Low            float64
       Close          float64
       Volume         object
       Market Cap     object
       dtype: object
```

Summary statistics:

```
In [3]: df.describe()
```

```
Out[3]:
```

	Open	High	Low	Close
count	1760.000000	1760.000000	1760.000000	1760.000000
mean	1479.574239	1536.991108	1418.251273	1485.747000
std	2950.457188	3092.950362	2780.859459	2959.265478
min	68.500000	74.560000	65.530000	68.430000
25%	274.730000	279.857500	268.640000	274.875000
50%	480.715000	495.195000	472.510000	482.810000
75%	870.085000	900.222500	831.682500	871.370000
max	19475.800000	20089.000000	18974.100000	19497.400000

Attributes description:

**Date:** date of observation

**Open:** Opening price on the given day

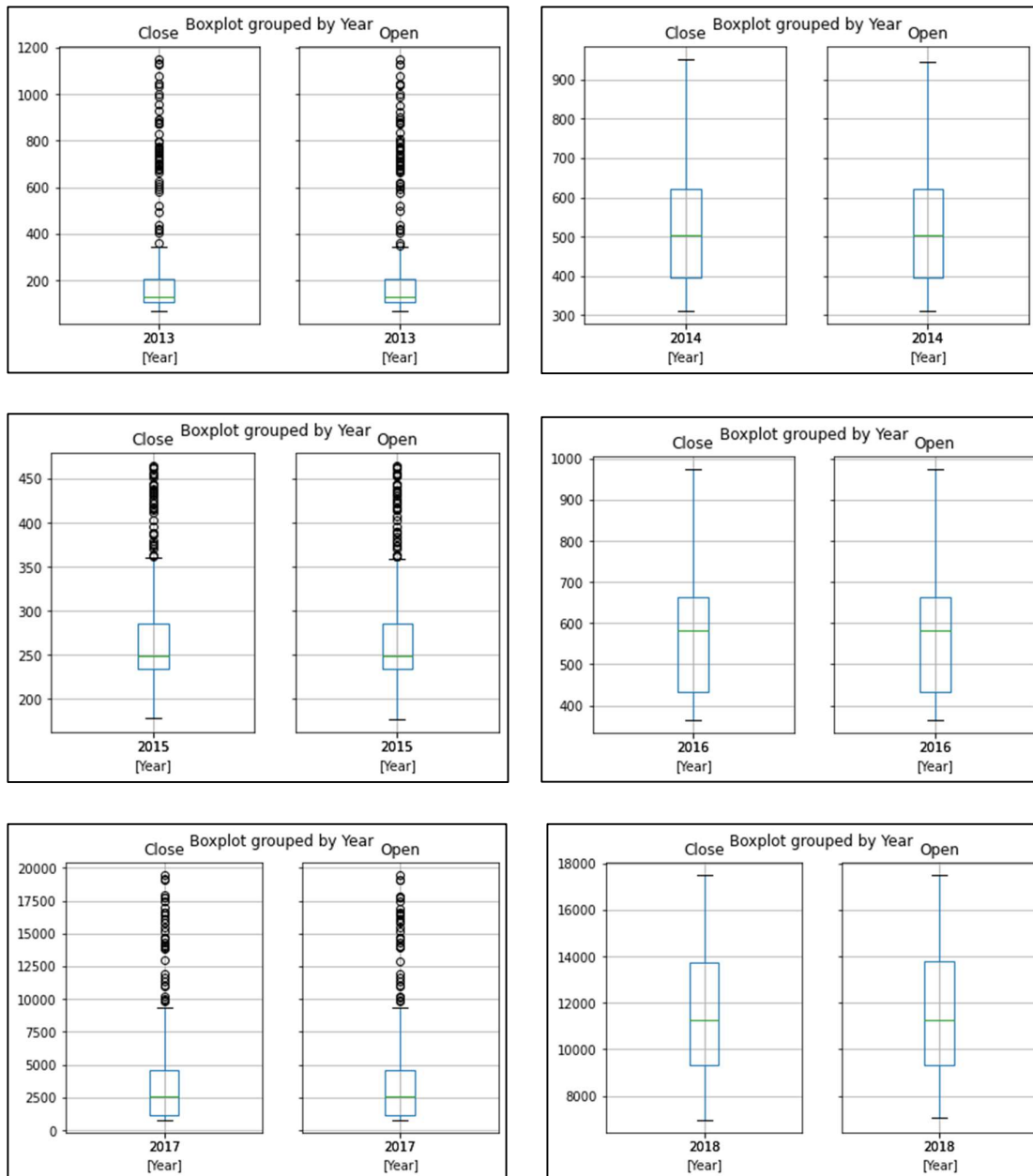
**High:** Highest price on the given day

**Low:** Lowest price on the given day

**Close:** Closing price on the given day  
**Volume:** Volume of transactions on the given day  
**Market\_Cap:** Market capitalization in USD

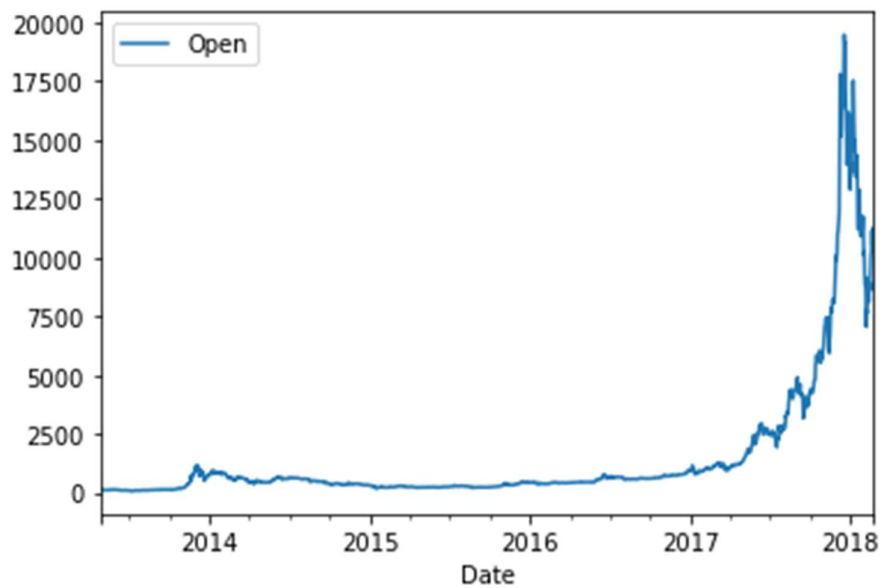
**Missing values:** There are no missing values in the dataset.  
The attribute Volume has 243 (14%) mismatched values

### Boxplots for Open and Close prices grouped by Year

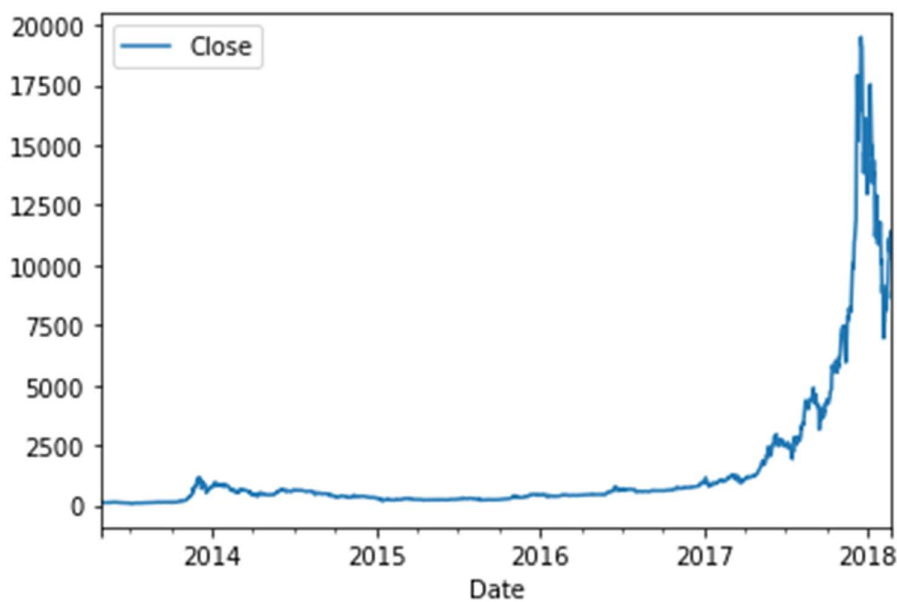


From the above boxplots, we can see that there are outliers present in the Open and Close prices for the years 2013, 2015 and 2017.

**Plot of Open prices from 28/04/2013**



**Plot of Close prices from 28/04/2013**



### **Data Cleaning**

The mismatched values in this dataset can be replaced by the mean value corresponding to the particular time intervals. The values of the Market\_Cap and Volume attributes must be converted to type float by removing the intermediate commas.

The values in the dataset seem to be of the same scale. There is no need for any transformation. All the attributes are relevant for the analysis. There seems to be no redundant attributes. Hence there is no need for dimensionality reduction.

## Ethereum\_price.csv

Number of columns: 7

Number of rows: 929

Attributes:

```
In [2]: df.dtypes
```

```
Out[2]: Date          datetime64[ns]
Open              float64
High              float64
Low               float64
Close             float64
Volume            object
Market_Cap        object
dtype: object
```

Summary statistics:

```
In [3]: df.describe()
```

```
Out[3]:
```

	Open	High	Low	Close
count	929.000000	929.000000	929.000000	929.000000
mean	146.897500	154.363413	139.114778	147.784327
std	263.218553	277.620377	247.137437	264.130846
min	0.431589	0.482988	0.420897	0.434829
25%	7.890000	8.270000	7.570000	7.910000
50%	12.050000	12.430000	11.720000	12.020000
75%	245.260000	257.000000	223.980000	245.990000
max	1397.480000	1432.880000	1290.600000	1396.420000

Attributes description:

**Date:** date of observation

**Open:** Opening price on the given day

**High:** Highest price on the given day

**Low:** Lowest price on the given day

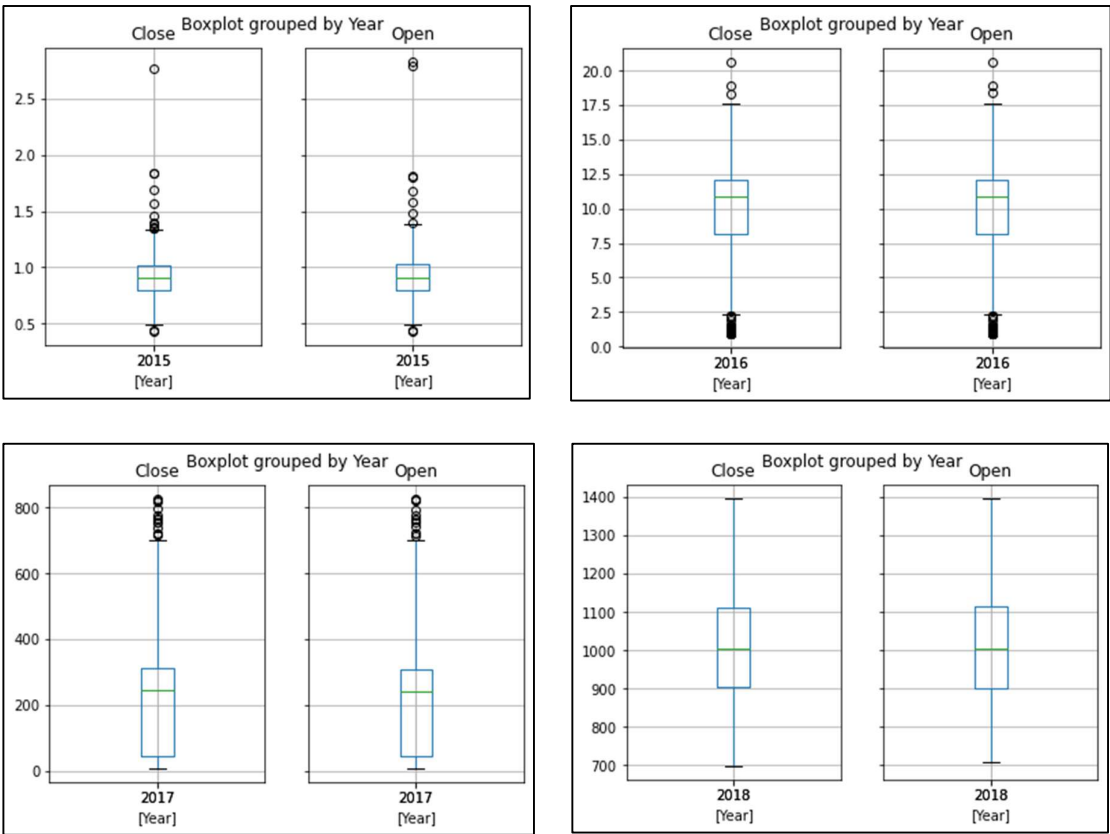
**Close:** Closing price on the given day

**Volume:** Volume of transactions on the given day

**Market\_Cap:** Market capitalization in USD

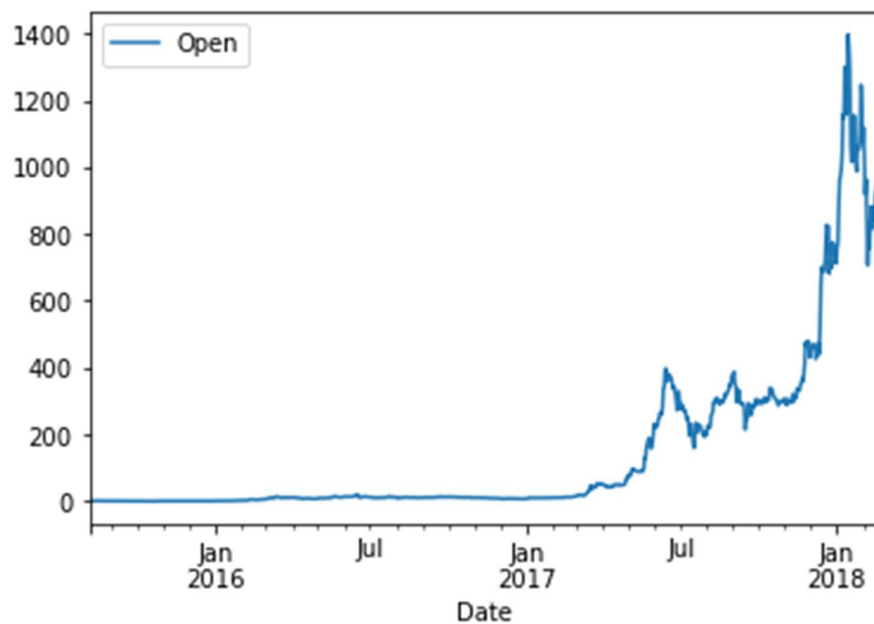
**Missing values:** There are no missing or mismatched values in the dataset.

Boxplots for Open and Close prices grouped by Year

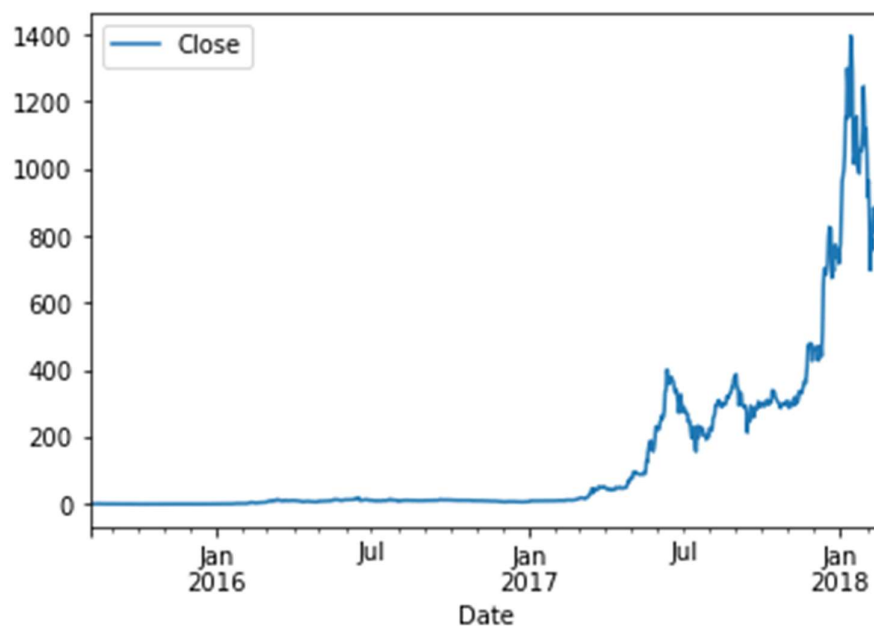


From the above boxplots, we can see that there are outliers present in the Open and Close prices for the years 2015, 2016 and 2017.

**Plot of Open prices from 07/08/2015**



**Plot of Close prices from 07/08/2015**



### **Data Cleaning**

The values of the Market\_Cap and Volume attributes must be converted to type float by removing the intermediate commas.

The values in the dataset seem to be of the same scale. There is no need for any transformation. All the attributes are relevant for the analysis. There seems to be no redundant attributes. Hence there is no need for dimensionality reduction.

## Ripple\_price.csv

Number of columns: 7

Number of rows: 1662

Attributes:

```
In [5]: df.dtypes

Out[5]: Date          datetime64[ns]
Open              float64
High              float64
Low               float64
Close             float64
Volume            object
Market Cap        object
dtype: object
```

Summary statistics:

```
In [6]: df.describe()

Out[6]:
```

	Open	High	Low	Close
count	1662.000000	1662.000000	1662.000000	1662.000000
mean	0.097768	0.105751	0.090019	0.098423
std	0.319687	0.352446	0.289041	0.320599
min	0.002809	0.003082	0.002802	0.002810
25%	0.006070	0.006192	0.005972	0.006070
50%	0.007953	0.008072	0.007819	0.007949
75%	0.019681	0.020643	0.018811	0.019830
max	3.360000	3.840000	3.120000	3.380000

Attributes description:

**Date:** date of observation

**Open:** Opening price on the given day

**High:** Highest price on the given day

**Low:** Lowest price on the given day

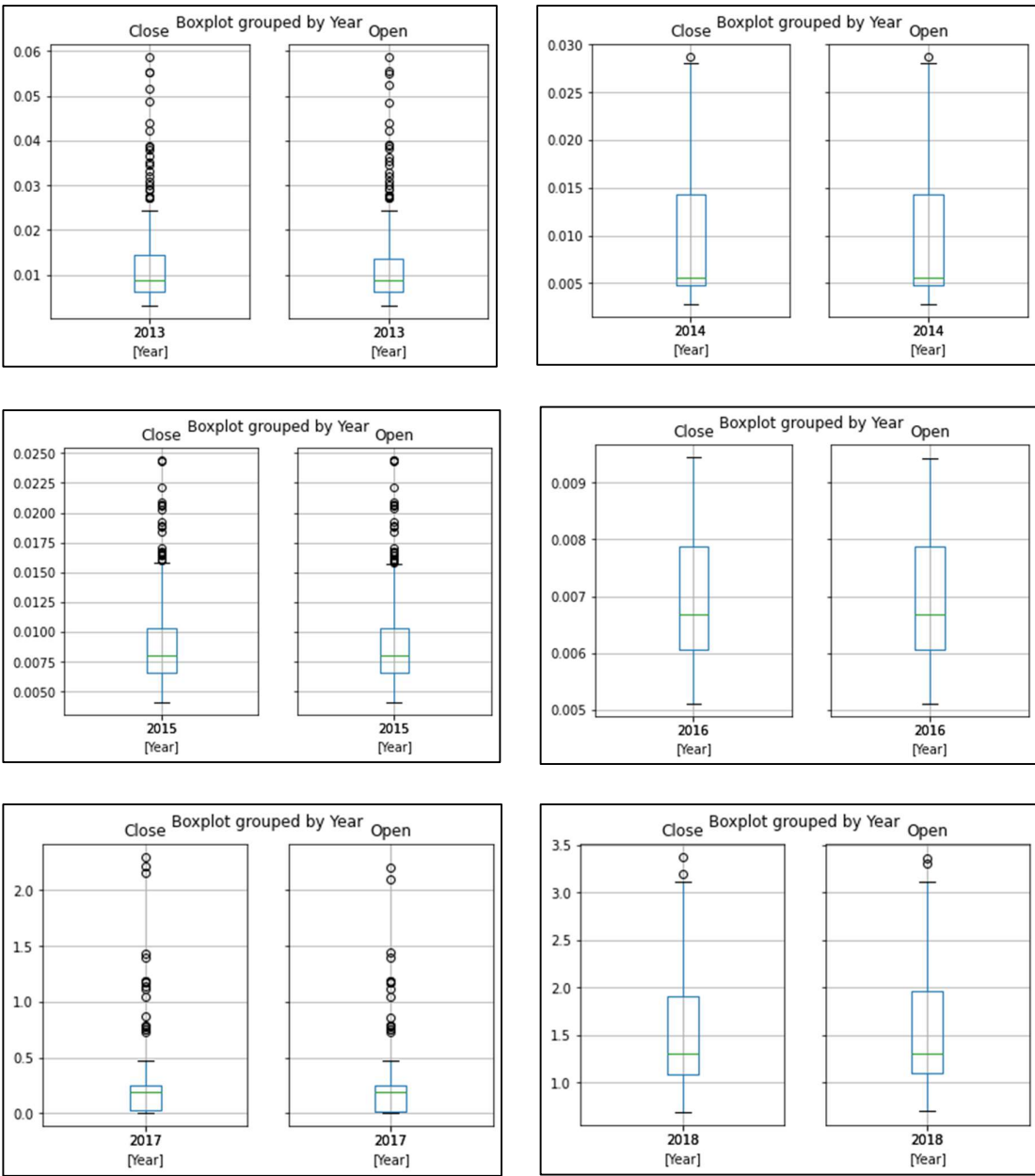
**Close:** Closing price on the given day

**Volume:** Volume of transactions on the given day

**Market\_Cap:** Market capitalization in USD

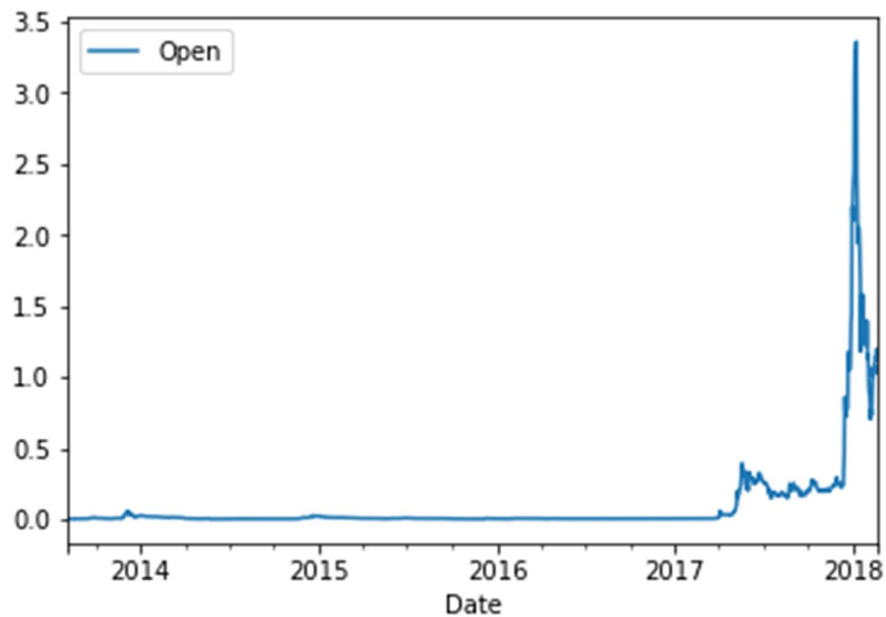
**Missing values:** There are no missing values in the dataset.  
The attribute Volume has 145 (9%) mismatched values.

Boxplots for Open and Close prices grouped by Year

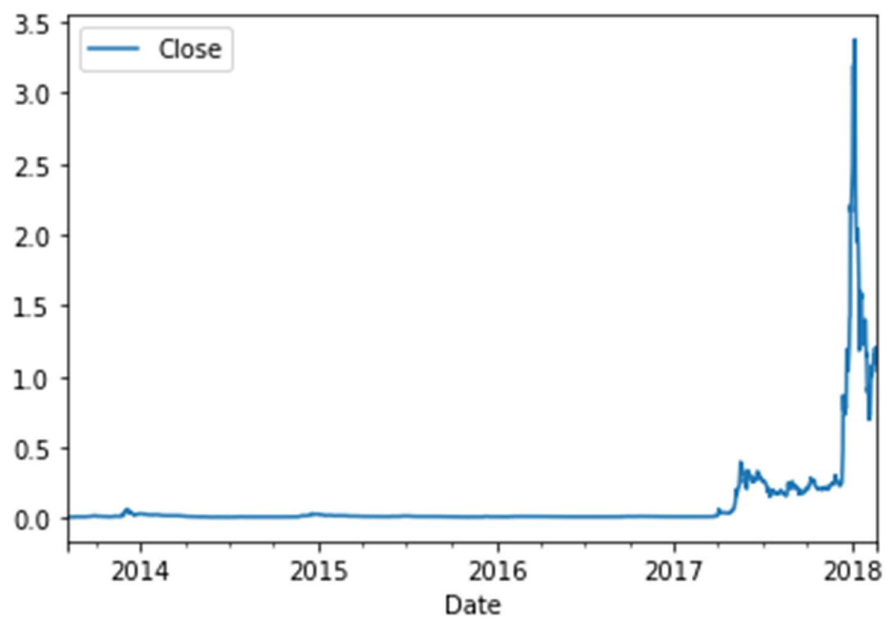




**Plot of Open prices from 04/08/2013**



**Plot of Close prices from 04/08/2013**



### **Data Cleaning**

The mismatched values in this dataset can be replaced by the mean value corresponding to the particular time intervals. The values of the Market\_Cap and Volume attributes must be converted to type float by removing the intermediate commas.

The values in the dataset seem to be of the same scale. There is no need for any transformation. All the attributes are relevant for the analysis. There seems to be no redundant attributes. Hence there is no need for dimensionality reduction.