Data Science Project Number 2

Mohammad Bagher Soroush

Project Master :

Mohammad Reza Momeni

Summer 2022

Introduction: the dataset that we have is about American financial market. Our goal is to finding noises in the dataset and analyze the Adj Close base on Date plot. Here's a quick look of our dataset (the first five and the last five lines):

	Index	Date	Open	High	Low	Close	Adj Close	Volume
0	NYA	12/31/1965	528.690002	528.690002	528.690002	528.690002	528.690002	0.0
1	NYA	1/3/1966	527.210022	527.210022	527.210022	527.210022	527.210022	0.0
2	NYA	1/4/1966	527.840027	527.840027	527.840027	527.840027	527.840027	0.0
3	NYA	1/5/1966	531.119995	531.119995	531.119995	531.119995	531.119995	0.0
4	NYA	1/6/1966	532.070007	532.070007	532.070007	532.070007	532.070007	0.0
112452	N100	5/27/2021	1241.119995	1251.910034	1241.119995	1247.069946	1247.069946	379696400.0
112453	N100	5/28/2021	1249.469971	1259.209961	1249.030029	1256.599976	1256.599976	160773400.0
112454	N100	5/31/2021	1256.079956	1258.880005	1248.140015	1248.930054	1248.930054	91173700.0
112455	N100	6/1/2021	1254.609985	1265.660034	1254.609985	1258.579956	1258.579956	155179900.0
112456	N100	6/2/2021	1258.489990	1263.709961	1258.239990	1263.619995	1263.619995	148465000.0

As you can see, our dataset has 112457 rows and 8 columns.

The columns include: index, date, open, high, low, close, adj close and volume.

First of all, we create a data frame and make the changes that have been requested:

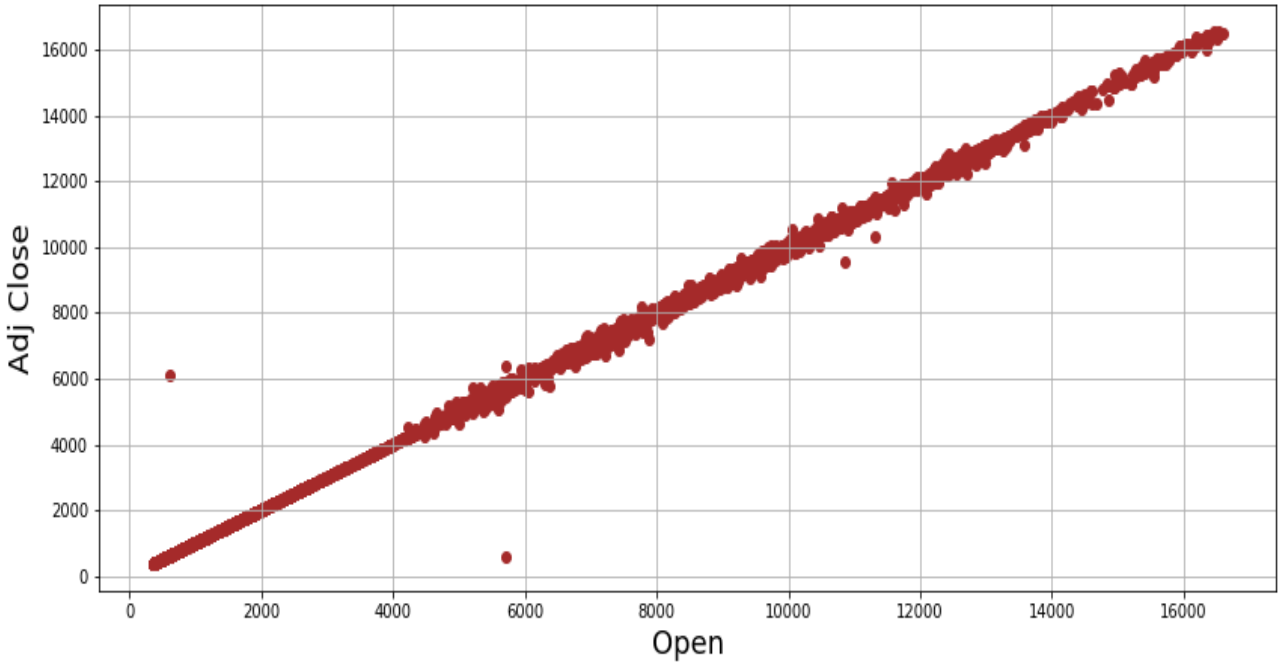
1. We have to sort our data frame buy index "NYA" which is New York stock exchange index.

2. We have to remove the decimal part of numbers in every numerical column.

3. We have to remove the "volume" column, because there's lot of zero value in it and it's impossible that we have this value for the total number of shares that have been bought or sold. Now we have to remove them and if we do that, we'll lost very much data and instead of that, we put the "volume" column away.

Now that we sort our data frame, it's time to draw "Adj Close" plots for finding noises.

Adj Close base on Open:

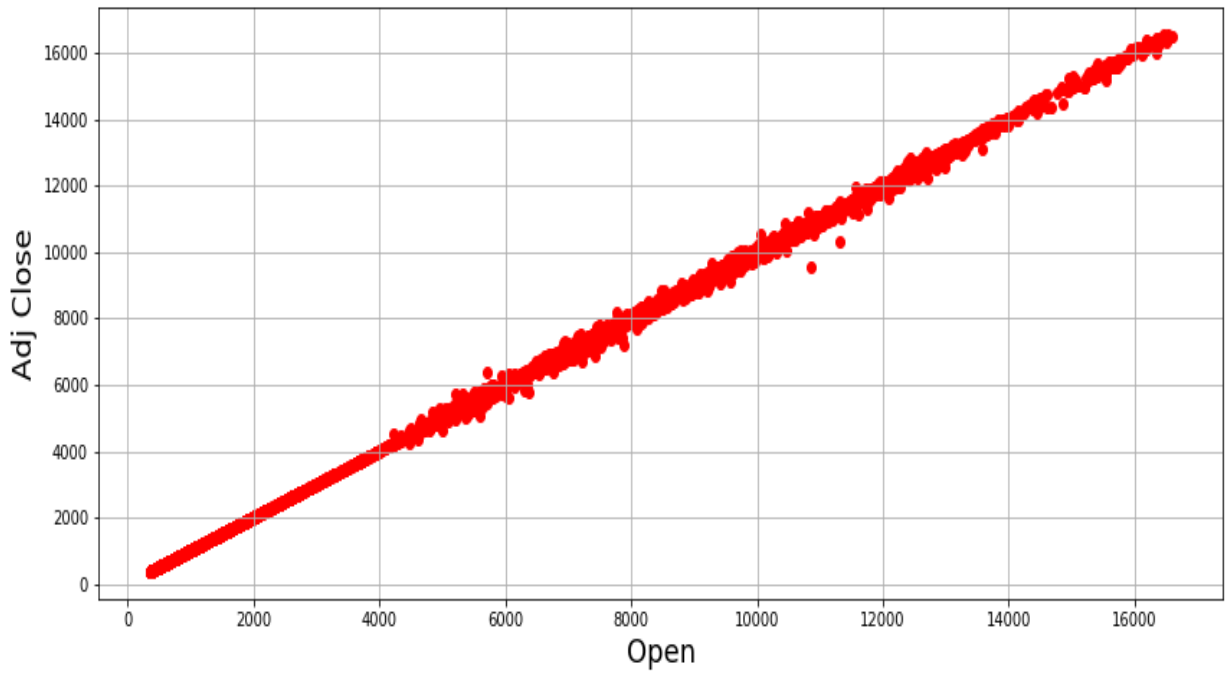


As you can see, there's two samples that away from the place of high density of data and they're seem to be noises and we have to remove the.

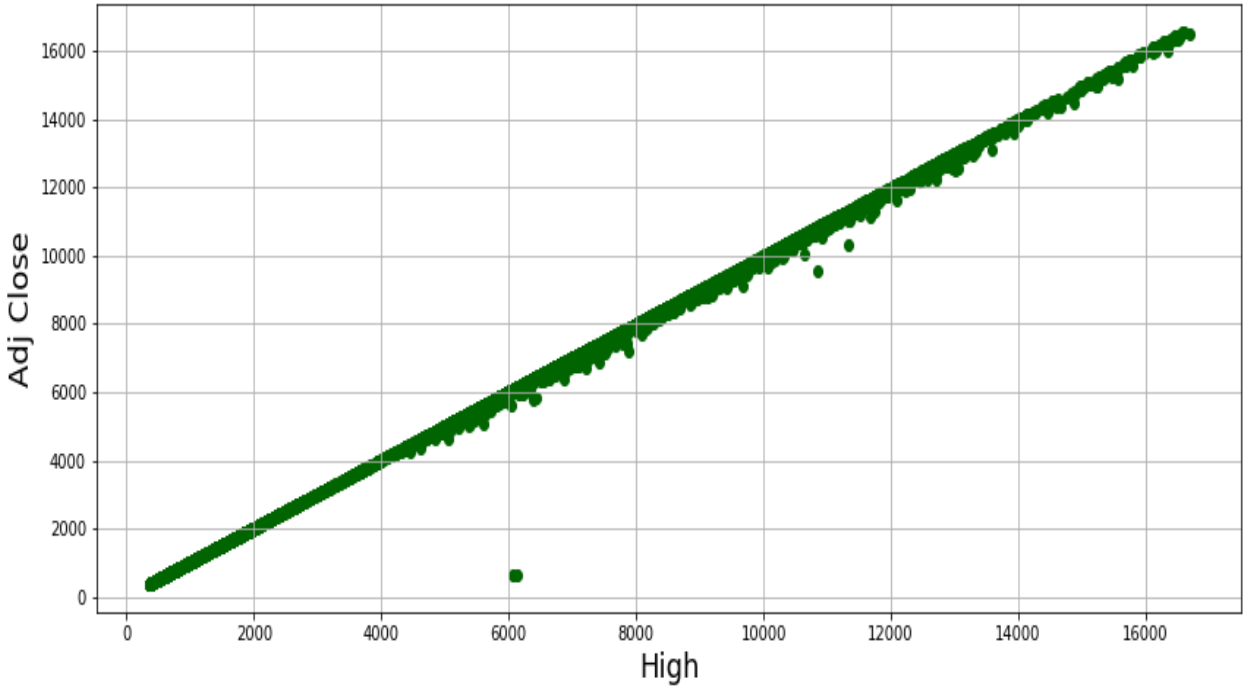
13943	NYA	5/24/2021	16375.0	16509.0	16375.0	16465.0	16465.0
13944	NYA	5/25/2021	16465.0	16526.0	16375.0	16390.0	16390.0
13945	NYA	5/26/2021	16390.0	16466.0	16388.0	16452.0	16452.0
13946	NYA	5/27/2021	16452.0	16546.0	16452.0	16532.0	16532.0
13947	NYA	5/28/2021	16532.0	16589.0	16532.0	16556.0	16556.0

13946 rows × 7 columns

(The number of row's has been reduced to 13946 after removing two noises)



Adj Close base on High:

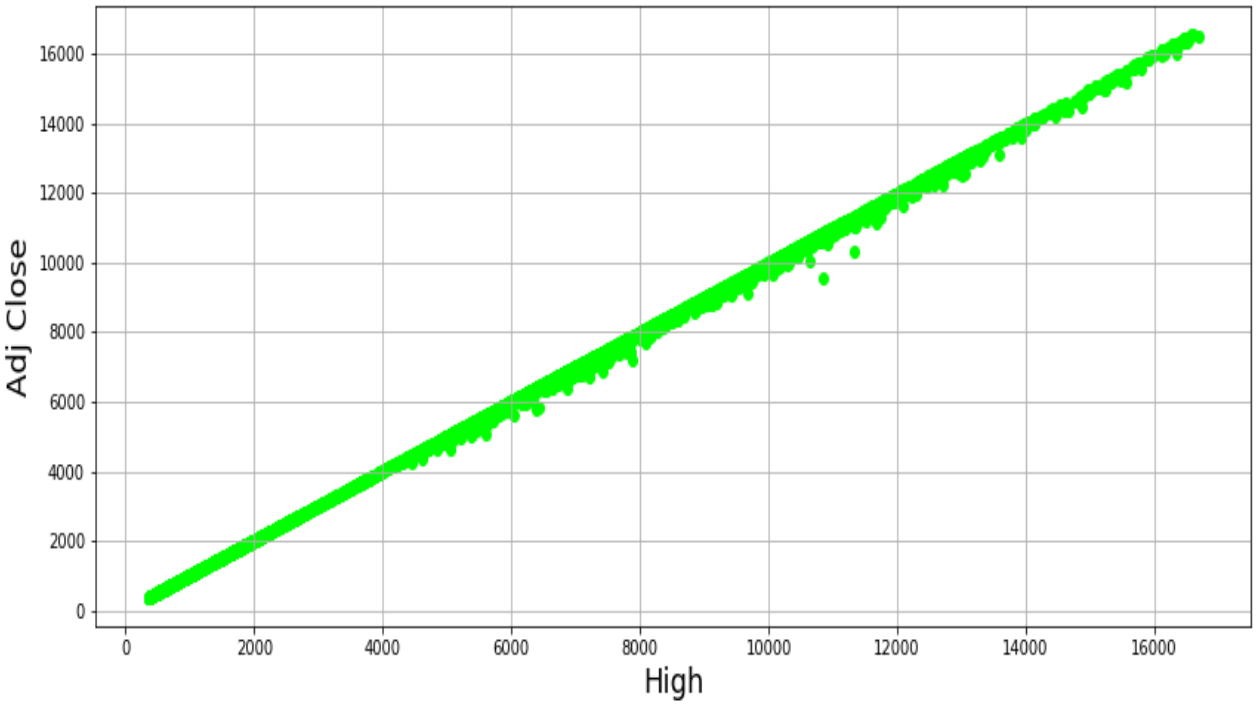


Regarding to the plot, there's some samples together far away from the line graph and seem to be noises and we have to remove them.

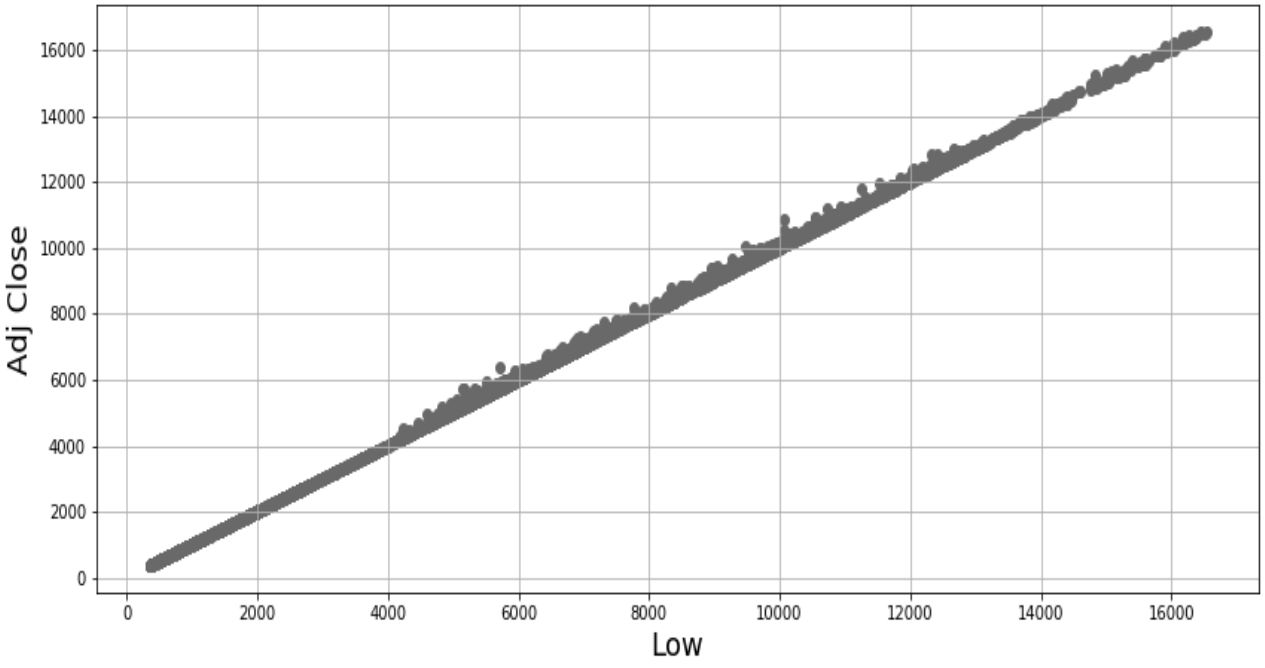
13943	NYA	5/24/2021	16375.0	16509.0	16375.0	16465.0	16465.0
13944	NYA	5/25/2021	16465.0	16526.0	16375.0	16390.0	16390.0
13945	NYA	5/26/2021	16390.0	16466.0	16388.0	16452.0	16452.0
13946	NYA	5/27/2021	16452.0	16546.0	16452.0	16532.0	16532.0
13947	NYA	5/28/2021	16532.0	16589.0	16532.0	16556.0	16556.0

13944 rows × 7 columns

(The number of rows has been reduced to 13944 after removing noises)

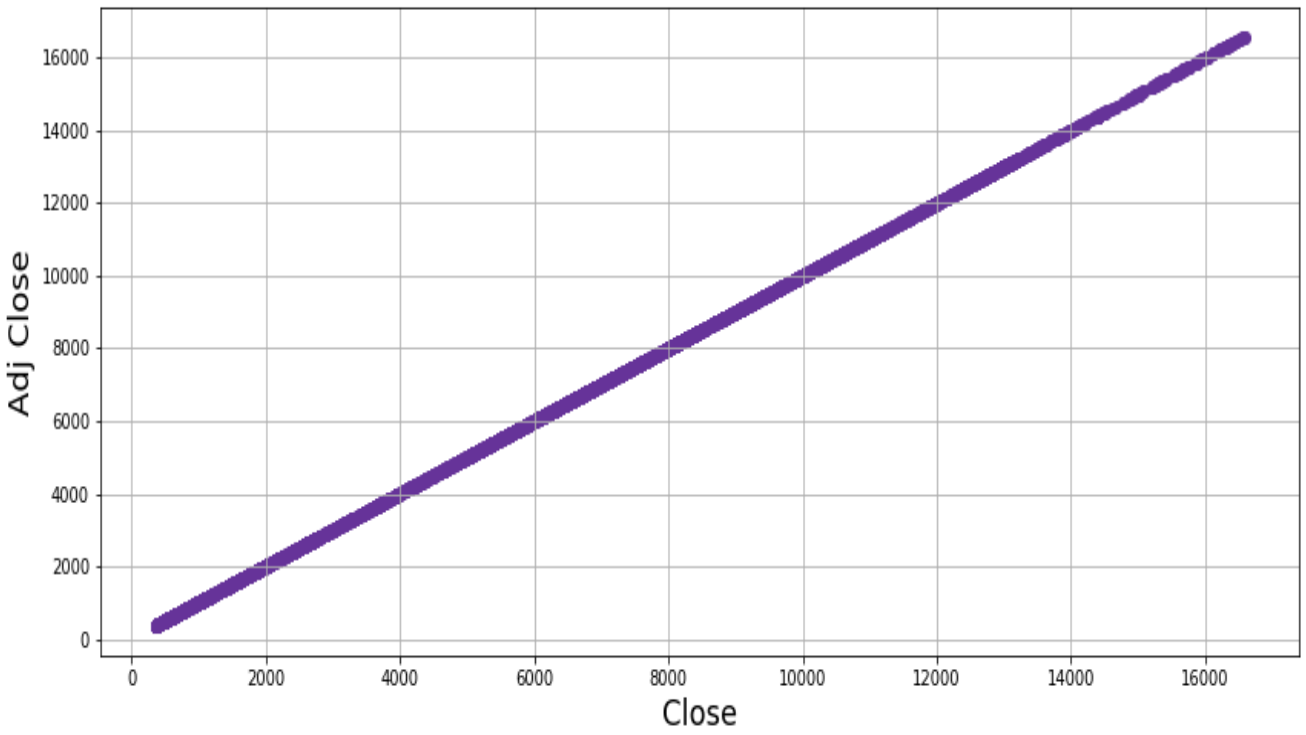


Adj Close base on Low:



There is no suspicious sample in this plot

Adj Close base on Close:



This line graph has normal process in this plot and there's no place to finding noises.

Now that we found noises and removed them, it's time to detecting missing values and clean our data frame.

```
Index          0
Date           0
Open           1
High           2
Low            3
Close          4
Adj Close     10
```

As you can see, we have some missing values in our columns except "Index", "Date" and now that we have no noises, we can easily drop them all.

13943	NYA	5/24/2021	16375.0	16509.0	16375.0	16465.0	16465.0
13944	NYA	5/25/2021	16465.0	16526.0	16375.0	16390.0	16390.0
13945	NYA	5/26/2021	16390.0	16466.0	16388.0	16452.0	16452.0
13946	NYA	5/27/2021	16452.0	16546.0	16452.0	16532.0	16532.0
13947	NYA	5/28/2021	16532.0	16589.0	16532.0	16556.0	16556.0

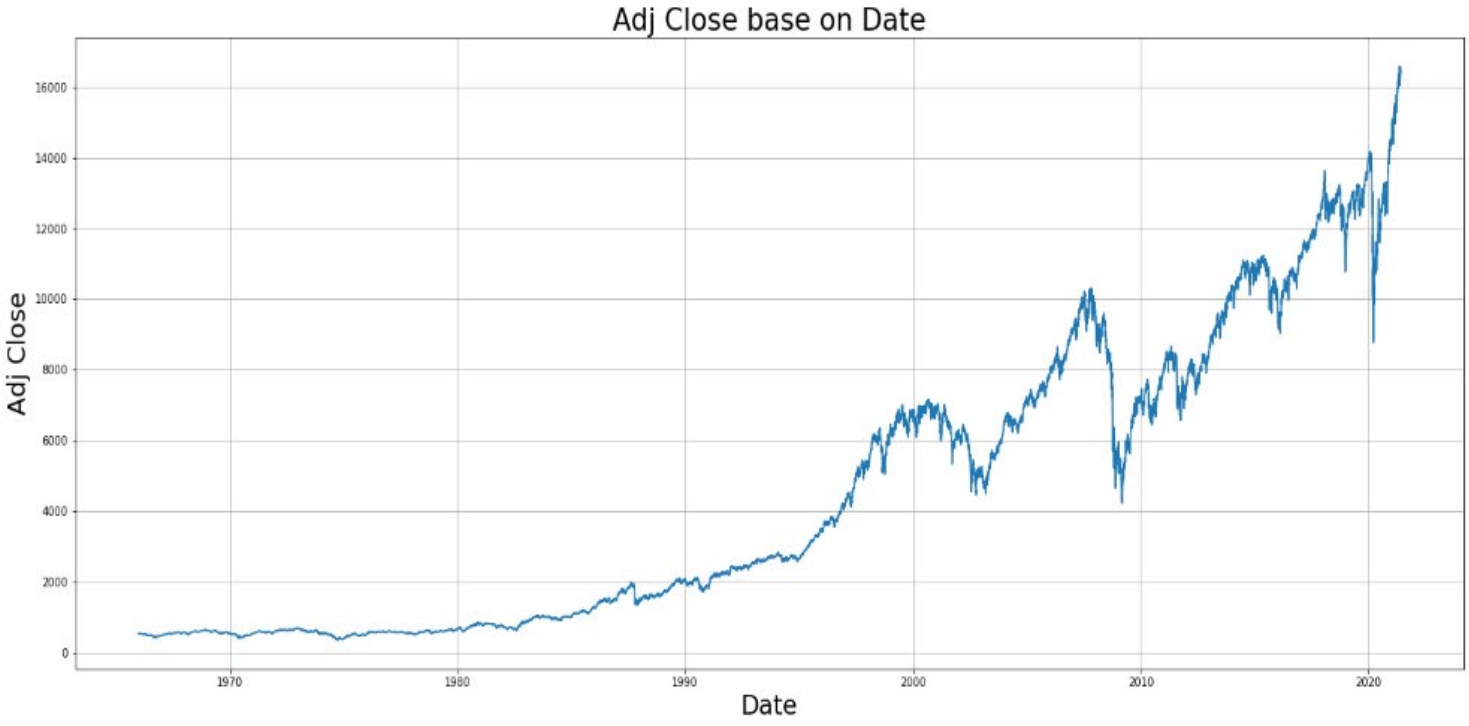
13928 rows × 7 columns

(The number of rows has been reduced 13928 and about 16 rows have been removed)

Now that our data frame is cleaned, we can draw our time series plots but before that, we have to change the format of date.

	Index	Date	Open	High	Low	Close	Adj Close
0	NYA	1965-12-31	529.0	529.0	529.0	529.0	529.0
1	NYA	1966-01-03	527.0	527.0	527.0	527.0	527.0
2	NYA	1966-01-04	528.0	528.0	528.0	528.0	528.0
3	NYA	1966-01-05	531.0	531.0	531.0	531.0	531.0
4	NYA	1966-01-06	532.0	532.0	532.0	532.0	532.0
13943	NYA	2021-05-24	16375.0	16509.0	16375.0	16465.0	16465.0
13944	NYA	2021-05-25	16465.0	16526.0	16375.0	16390.0	16390.0
13945	NYA	2021-05-26	16390.0	16466.0	16388.0	16452.0	16452.0
13946	NYA	2021-05-27	16452.0	16546.0	16452.0	16532.0	16532.0
13947	NYA	2021-05-28	16532.0	16589.0	16532.0	16556.0	16556.0

Adj Close base on Date:



regarding to the plot, we have a positive slop between 1970 and 2000. But this slop become reversed sins 2001 to 2003 and back to positive again till 2008. After that year, our stack's Adj Close has fallen sharply. We also have such situation in 2020 but sharper. The third positive slop starts at 2009 till 2020 but between 2014 and 2016 we have a negative slop in general.

Result: analyzing graph behavior and cause of the positive or negative slop of graph in plot required more information and study, but we can say 2001, 2003, 2008, 2009, 2015, 2016 and 2020 were impactive break points for our plot and important health, political and economic issues must have been involved in it.

Project code and description of its steps: