تمرین جلسه 9

امیر محمد استادکار 95

```
#######  amir mohammad ostadkar   ##95
######    jalaseh 9  data
#####  step 1 =  flout change to ++++++  int
#####  step 2 = index certion ====  NYI
#####  step 3 ====  datatime frame ----->   plot  >>> x= data    y = adj close
#####   step 4 =   delete noise with reason
#####    step 5    find missing value and fillna or dropna
```

مرحله صفر

Import data and lib

```
data = pd.read_csv('Market.csv')
df = pd.DataFrame(data)
df
```

	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

Describe data

```
df.describe().transpose()
```

	count	mean	std	min	25%	50%	75%	max
Open	110253.0	7.658562e+03	9.011456e+03	54.869999	1855.060059	5194.399902	1.013430e+04	6.877506e+04
High	110252.0	7.704538e+03	9.066605e+03	54.869999	1864.687470	5226.750000	1.020783e+04	6.940375e+04
Low	110251.0	7.608130e+03	8.954537e+03	54.869999	1844.015015	5154.299805	1.006037e+04	6.851699e+04
Close	110250.0	7.657741e+03	9.011556e+03	54.869999	1855.347473	5194.889892	1.013487e+04	6.877506e+04
Adj Close	110244.0	7.657983e+03	9.011724e+03	54.869999	1855.057556	5195.699951	1.013551e+04	6.877506e+04
Volume	110253.0	1.273975e+09	4.315783e+09	0.000000	0.000000	432900.000000	1.734314e+08	9.440374e+10

```
# اطلاعات دارای میسینگ ولو میباشد
#  دارای اعشار میباشد
```

Step 1 int number

```
df2 = df.round(0)
df2
```

	Index	Date	Open	High	Low	Close	Adj Close	Volume
0	NYA	12/31/1965	529.0	529.0	529.0	529.0	529.0	0.0
1	NYA	1/3/1966	527.0	527.0	527.0	527.0	527.0	0.0
2	NYA	1/4/1966	528.0	528.0	528.0	528.0	528.0	0.0
3	NYA	1/5/1966	531.0	531.0	531.0	531.0	531.0	0.0
4	NYA	1/6/1966	532.0	532.0	532.0	532.0	532.0	0.0
...	...	...	...	...	...	...	...	...
112452	N100	5/27/2021	1241.0	1252.0	1241.0	1247.0	1247.0	379696400.0
112453	N100	5/28/2021	1249.0	1259.0	1249.0	1257.0	1257.0	160773400.0
112454	N100	5/31/2021	1256.0	1259.0	1248.0	1249.0	1249.0	91173700.0
112455	N100	6/1/2021	1255.0	1266.0	1255.0	1259.0	1259.0	155179900.0
112456	N100	6/2/2021	1258.0	1264.0	1258.0	1264.0	1264.0	148465000.0

112457 rows × 8 columns

Step 2 === index = NYA

```
##### step 2      certion index ======NYA
```

```
df3 = df2[df2['Index']=='NYA']
df3
```

	Index	Date	Open	High	Low	Close	Adj Close	Volume
0	NYA	12/31/1965	529.0	529.0	529.0	529.0	529.0	0.000000e+00
1	NYA	1/3/1966	527.0	527.0	527.0	527.0	527.0	0.000000e+00
2	NYA	1/4/1966	528.0	528.0	528.0	528.0	528.0	0.000000e+00
3	NYA	1/5/1966	531.0	531.0	531.0	531.0	531.0	0.000000e+00
4	NYA	1/6/1966	532.0	532.0	532.0	532.0	532.0	0.000000e+00
...	...	...	...	...	...	...	...	...
13943	NYA	5/24/2021	16375.0	16509.0	16375.0	16465.0	16465.0	2.947400e+09
13944	NYA	5/25/2021	16465.0	16526.0	16375.0	16390.0	16390.0	3.420870e+09
13945	NYA	5/26/2021	16390.0	16466.0	16388.0	16452.0	16452.0	3.674490e+09
13946	NYA	5/27/2021	16452.0	16546.0	16452.0	16532.0	16532.0	5.201110e+09
13947	NYA	5/28/2021	16532.0	16589.0	16532.0	16556.0	16556.0	4.199270e+09

13948 rows × 8 columns

```
df3.describe().transpose()
```

	count	mean	std	min	25%	50%	75%	max
Open	13947.0	4.452145e+03	4.074833e+03	348.0	655.0	2632.0	7.339500e+03	1.659000e+04
High	13946.0	4.469313e+03	4.094960e+03	348.0	655.0	2632.0	7.376500e+03	1.668600e+04
Low	13945.0	4.434261e+03	4.052815e+03	348.0	655.0	2632.0	7.278000e+03	1.653200e+04
Close	13944.0	4.453026e+03	4.075485e+03	348.0	655.0	2632.0	7.339750e+03	1.659000e+04
Adj Close	13938.0	4.455094e+03	4.075458e+03	348.0	656.0	2633.0	7.342750e+03	1.659000e+04
Volume	13947.0	1.215565e+09	1.834155e+09	0.0	0.0	0.0	2.681975e+09	1.145623e+10

missing value

Step 3 == data === datatime

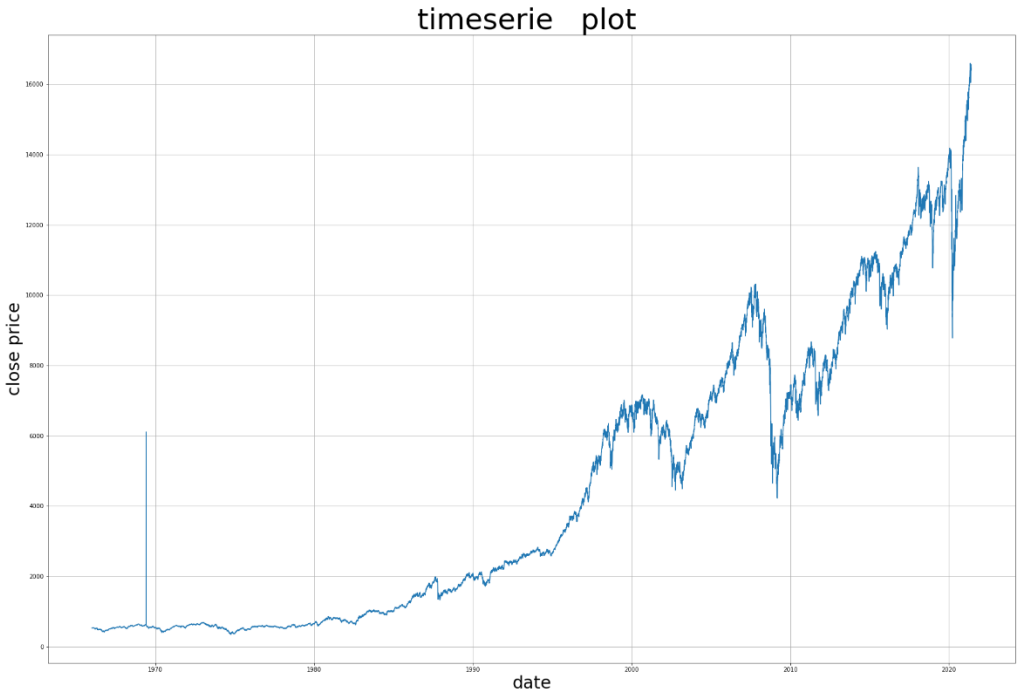
And plot

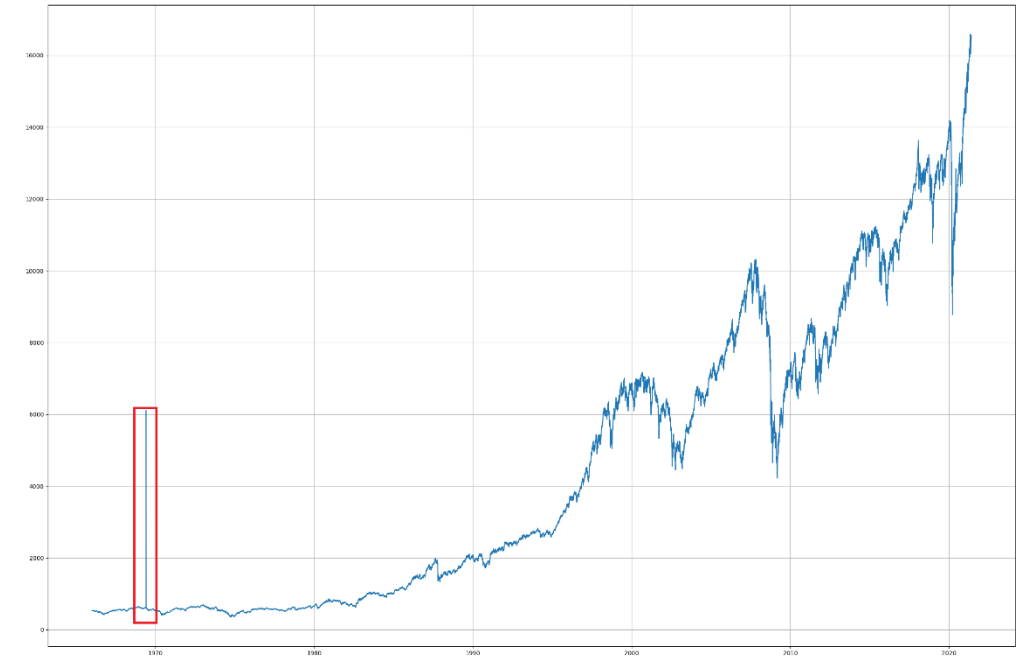
```
#######  STEP 3    plot  x= date    y  = adj close
```

```
###########  info from df3
```

```
df3.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 13948 entries, 0 to 13947
Data columns (total 8 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Index      13948 non-null  object
 1   Date       13948 non-null  object
 2   Open       13947 non-null  float64
 3   High       13946 non-null  float64
 4   Low        13945 non-null  float64
 5   Close      13944 non-null  float64
 6   Adj Close  13938 non-null  float64
 7   Volume     13947 non-null  float64
dtypes: float64(6), object(2)
memory usage: 980.7+ KB
```


timeserie plot
close price
date
0
2000
4000
6000
8000
10000
12000
14000
16000
1970
1980
1990
2000
2010
2020



```
:
```

```
: ################   شناسایی نویز در محدوده تا سال هزار نهصد هفتاد
  ############   و در محدوده قیمتی 6000
```

```
: #########   نویز شناسایی شد
```

```
: #########   این مقدار عدد نبوده   دلیل === به دلیل فاصله زیاد از تاریخ 1970 تا 2010
```

```
: #########   دلیل دو      در این محدوده تا سال 2010 میانگین کمتر از 6000  بوده ولی داده بالای 6000 نشان میدهد
```

شناسایی داده های گمشده

```
print('\show missing value in Df:\n\n',df3.isnull().sum())
```

```
\show missing value in Df:

 Index         0
Date           0
Open           1
High           2
Low            3
Close          4
Adj Close     10
Volume         1
day            0
month          0
year           0
dtype: int64
```

```
#######  نمایش تعداد میسینگ ولو و ادرس انها
for i in range(len(df3.index)):
    if (df.iloc[i].isnull().sum()) >0:
        print(('total nan ',i),df.iloc[i].isnull().sum())
```

```
('total nan ', 102) 1
('total nan ', 104) 1
('total nan ', 154) 1
('total nan ', 170) 1
('total nan ', 190) 1
('total nan ', 231) 1
('total nan ', 257) 1
('total nan ', 282) 1
('total nan ', 289) 6
('total nan ', 307) 1
('total nan ', 333) 1
('total nan ', 353) 1
('total nan ', 464) 1
('total nan ', 635) 1
('total nan ', 700) 1
('total nan ', 800) 1
```

حذف نویز

```
##### ایجاد ستون جدید در دیتا فریم
df3['day'] = df3['Date'].dt.day
df3['month'] = df3['Date'].dt.month
df3['year'] = df3['Date'].dt.year
df3
```

```
#   شناسایی نویز با محدود کردن دیتا فریم
df500 = df3[(df3['year']<1986)&(df3['Adj Close']>4000)]
df500
```

	Index	Date	Open	High	Low	Close	Adj Close	Volume	day	month	year
831	NYA	1969-05-29 00:00:00+00:00	612.0	612.0	612.0	612.0	6111.0	0.0	29	5	1969

```
###############  نویز    ادرس و اندکس دقیق
```

```
####  drop noise
df6 = df3.drop(df3.index[831])
df6
```

اطلاعات کامل در فایل ژوپیتر