بررسی دیتاست بازار مالی

Milad Hashemi

داده هایی به شرح زیر۱۱۲۴۵۷× ۸دریافت شده است:

	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

112457 rows × 8 columns

بعد از بررسی , اطلاعات اولیه به شرح زیر از داده دریافت شده است:

	Index	Date	Open	High	Low	Close	Adj Close	Volume
count	112457	112457	110253.000000	110252.000000	110251.000000	110250.000000	110244.000000	1.102530e+05
unique	14	14731	NaN	NaN	NaN	NaN	NaN	NaN
top	N225	11/3/2017	NaN	NaN	NaN	NaN	NaN	NaN
freq	14500	14	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	NaN	7658.561932	7704.538416	7608.129749	7657.740822	7657.982529	1.273975e+09
std	NaN	NaN	9011.455529	9066.605458	8954.536718	9011.555549	9011.723572	4.315783e+09
min	NaN	NaN	54.869999	54.869999	54.869999	54.869999	54.869999	0.000000e+00
25%	NaN	NaN	1855.060059	1864.687470	1844.015015	1855.347473	1855.057556	0.000000e+00
50%	NaN	NaN	5194.399902	5226.750000	5154.299805	5194.889892	5195.699951	4.329000e+05
75%	NaN	NaN	10134.299810	10207.827635	10060.369630	10134.867430	10135.512452	1.734314e+08
max	NaN	NaN	68775.062500	69403.750000	68516.992190	68775.062500	68775.062500	9.440374e+10

از انجایی که کارفرما فقط ایندکس NYA رو مدنظر داشت برای همین این ایندکس از کل دیتاست جدا شد:

	Index	Date	Open	High	Low	Close	Adj Close	Volume
0	NYA	12/31/1965	528.690002	528.690002	528.690002	528.690002	528.690002	0.000000e+00
1	NYA	1/3/1966	527.210022	527.210022	527.210022	527.210022	527.210022	0.000000e+00
2	NYA	1/4/1966	527.840027	527.840027	527.840027	527.840027	527.840027	0.000000e+00
3	NYA	1/5/1966	531.119995	531.119995	531.119995	531.119995	531.119995	0.000000e+00
4	NYA	1/6/1966	532.070007	532.070007	532.070007	532.070007	532.070007	0.000000e+00
...	...	...	...	...	...	...	...	...
13943	NYA	5/24/2021	16375.000000	16508.519530	16375.000000	16464.689450	16464.689450	2.947400e+09
13944	NYA	5/25/2021	16464.689450	16525.810550	16375.150390	16390.189450	16390.189450	3.420870e+09
13945	NYA	5/26/2021	16390.189450	16466.339840	16388.320310	16451.960940	16451.960940	3.674490e+09
13946	NYA	5/27/2021	16451.960940	16546.359380	16451.960940	16531.949220	16531.949220	5.201110e+09
13947	NYA	5/28/2021	16531.949220	16588.689450	16531.949220	16555.660160	16555.660160	4.199270e+09

13948 rows × 8 columns

بعد از بررسی , اطلاعات اولیه به شرح زیر از داده دریافت شده است که missing value از یک تا ده تا در این دیتاست وجود دارد:

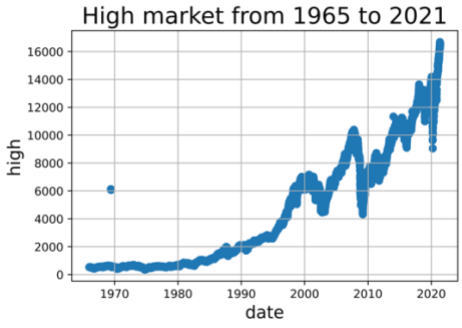
	Index	Date	Open	High	Low	Close	Adj Close	Volume
count	13948	13948	13947.000000	13946.000000	13945.000000	13944.000000	13938.000000	1.394700e+04
unique	1	13948	NaN	NaN	NaN	NaN	NaN	NaN
top	NYA	12/31/1965	NaN	NaN	NaN	NaN	NaN	NaN
freq	13948	1	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	NaN	4452.147406	4469.312526	4434.262223	4453.026486	4455.094446	1.215565e+09
std	NaN	NaN	4074.835507	4094.956718	4052.815490	4075.483921	4075.456765	1.834155e+09
min	NaN	NaN	347.769989	347.769989	347.769989	347.769989	347.769989	0.000000e+00
25%	NaN	NaN	654.989990	655.150024	655.039978	655.122513	655.807525	0.000000e+00
50%	NaN	NaN	2631.909912	2632.280029	2631.909912	2632.015014	2633.015014	0.000000e+00
75%	NaN	NaN	7339.489990	7376.315063	7277.509766	7339.397583	7342.787598	2.681975e+09
max	NaN	NaN	16590.429690	16685.890630	16531.949220	16590.429690	16590.429690	1.145623e+10

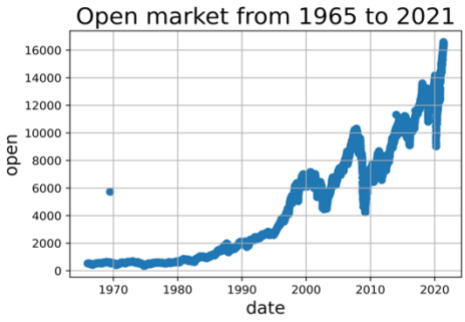
کل دیتاست بر اساس تاریخ (قدیم به جدید) مرتب شد تا نمودار واضح و با کیفیتی رسم شود:

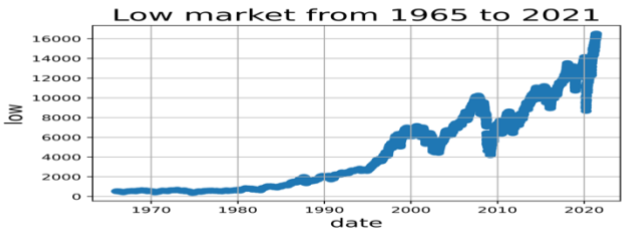
	Index	Date	Open	High	Low	Close	Adj Close	Volume
0	NYA	1965-12-31	528.690002	528.690002	528.690002	528.690002	528.690002	0.000000e+00
1	NYA	1966-01-03	527.210022	527.210022	527.210022	527.210022	527.210022	0.000000e+00
2	NYA	1966-01-04	527.840027	527.840027	527.840027	527.840027	527.840027	0.000000e+00
3	NYA	1966-01-05	531.119995	531.119995	531.119995	531.119995	531.119995	0.000000e+00
4	NYA	1966-01-06	532.070007	532.070007	532.070007	532.070007	532.070007	0.000000e+00
...	...	...	...	...	...	...	...	...
13943	NYA	2021-05-24	16375.000000	16508.519530	16375.000000	16464.689450	16464.689450	2.947400e+09
13944	NYA	2021-05-25	16464.689450	16525.810550	16375.150390	16390.189450	16390.189450	3.420870e+09
13945	NYA	2021-05-26	16390.189450	16466.339840	16388.320310	16451.960940	16451.960940	3.674490e+09
13946	NYA	2021-05-27	16451.960940	16546.359380	16451.960940	16531.949220	16531.949220	5.201110e+09
13947	NYA	2021-05-28	16531.949220	16588.689450	16531.949220	16555.660160	16555.660160	4.199270e+09

13948 rows × 8 columns

برای هر فیچر به طور جدا گانه scatter رسم شد تا داده های نویز شناسایی شود که در اینجا طبق پلات های زیر ما در فیچر های (Open,High,Adj close) داده نویز داشتیم:







داده های نویز در دیتاست پیدا شد و به صاحب دیتاست گزارش داده شد تا در صورت اصلاح, داده یی حذف نشود ولی با صحبت با اقای مومنی تصمیم به حذف نویز گرفته شد نویز ها یافت شده به شرح زیر است:

	Index	Date	Open	High	Low	Close	Adj Close	Volume
852	NYA	1969-06-30	5722.359985	572.359985	572.359985	572.359985	572.359985	0.0
711	NYA	1968-11-29	647.849976	647.849976	647.849976	647.849976	647.849976	0.0
712	NYA	1968-12-02	646.479980	646.479980	646.479980	646.479980	646.479980	0.0
713	NYA	1968-12-03	645.950012	645.950012	645.950012	645.950012	645.950012	0.0
715	NYA	1968-12-06	645.309998	645.309998	645.309998	645.309998	645.309998	0.0

	Index	Date	Open	High	Low	Close	Adj Close	Volume
829	NYA	1969-05-27	612.219971	6124.219971	612.219971	612.219971	612.219971	0.0
833	NYA	1969-06-03	606.830017	6065.830017	606.830017	606.830017	606.830017	0.0
711	NYA	1968-11-29	647.849976	647.849976	647.849976	647.849976	647.849976	0.0
712	NYA	1968-12-02	646.479980	646.479980	646.479980	646.479980	646.479980	0.0
713	NYA	1968-12-03	645.950012	645.950012	645.950012	645.950012	645.950012	0.0

	Index	Date	Open	High	Low	Close	Adj Close	Volume
831	NYA	1969-05-29	611.900024	611.900024	611.900024	611.900024	6110.900024	0.0
1745	NYA	1973-01-11	692.369995	692.369995	692.369995	692.369995	692.369995	0.0
1741	NYA	1973-01-05	691.200012	691.200012	691.200012	691.200012	691.200012	0.0
1742	NYA	1973-01-08	690.989990	690.989990	690.989990	690.989990	690.989990	0.0
1739	NYA	1973-01-03	690.359985	690.359985	690.359985	690.359985	690.359985	0.0

در ایندکس های ۸۵۲ و۸۲۹ و۸۳۱ داده نویز وجود داشته و در سطر مربوطه هیچ missing values وجود ندارد! بعد ار حذف نویز ها ما جدولی ار دیتاست ۱۳۹۴۴×۸ به شرح زیر داریم:

	Index	Date	Open	High	Low	Close	Adj Close	Volume
0	NYA	1965-12-31	528.690002	528.690002	528.690002	528.690002	528.690002	0.000000e+00
1	NYA	1966-01-03	527.210022	527.210022	527.210022	527.210022	527.210022	0.000000e+00
2	NYA	1966-01-04	527.840027	527.840027	527.840027	527.840027	527.840027	0.000000e+00
3	NYA	1966-01-05	531.119995	531.119995	531.119995	531.119995	531.119995	0.000000e+00
4	NYA	1966-01-06	532.070007	532.070007	532.070007	532.070007	532.070007	0.000000e+00
...	...	...	...	...	...	...	...	...
13943	NYA	2021-05-24	16375.000000	16508.519530	16375.000000	16464.689450	16464.689450	2.947400e+09
13944	NYA	2021-05-25	16464.689450	16525.810550	16375.150390	16390.189450	16390.189450	3.420870e+09
13945	NYA	2021-05-26	16390.189450	16466.339840	16388.320310	16451.960940	16451.960940	3.674490e+09
13946	NYA	2021-05-27	16451.960940	16546.359380	16451.960940	16531.949220	16531.949220	5.201110e+09
13947	NYA	2021-05-28	16531.949220	16588.689450	16531.949220	16555.660160	16555.660160	4.199270e+09

13944 rows × 8 columns

با بررسی داده طبق دیتاست زیر از یک تا ۱۰, missing values در جدول زیر مشاهده شده است:

	Open	High	Low	Close	Adj Close	Volume
count	13943.000000	13942.000000	13941.000000	13940.000000	13934.000000	1.394300e+04
mean	4452.882920	4469.635502	4435.362126	4454.131852	4455.806235	1.215914e+09
std	4075.016076	4095.234524	4052.876577	4075.546081	4075.624383	1.834302e+09
min	347.769989	347.769989	347.769989	347.769989	347.769989	0.000000e+00
25%	655.150024	655.230011	655.469971	655.544998	656.017487	0.000000e+00
50%	2632.120117	2632.280029	2632.439941	2632.439941	2633.704956	0.000000e+00
75%	7342.455078	7377.210083	7278.419922	7342.082519	7345.524902	2.682940e+09
max	16590.429690	16685.890630	16531.949220	16590.429690	16590.429690	1.145623e+10

بعد از صحبت با کارفرما و بررسی, تعداد missing values در هر سطر و ایندکس آن که به شرح زیر است تصمیم به حذف آن گرفته شد:

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

```
there is 1 missing value in row 102
there is 1 missing value in row 104
there is 1 missing value in row 154
there is 1 missing value in row 170
there is 1 missing value in row 190
there is 1 missing value in row 231
there is 1 missing value in row 257
there is 1 missing value in row 282
there are 6 missing values in row 289
there is 1 missing value in row 307
there is 1 missing value in row 333
there is 1 missing value in row 353
there is 1 missing value in row 464
there is 1 missing value in row 635
there is 1 missing value in row 700
there is 1 missing value in row 800
```

بعد از حذف missing values دیتاست۱۳۹۲۸×۸ مربوطه به شرح زیر است:

	Index	Date	Open	High	Low	Close	Adj Close	Volume
0	NYA	1965-12-31	528.690002	528.690002	528.690002	528.690002	528.690002	0.000000e+00
1	NYA	1966-01-03	527.210022	527.210022	527.210022	527.210022	527.210022	0.000000e+00
2	NYA	1966-01-04	527.840027	527.840027	527.840027	527.840027	527.840027	0.000000e+00
3	NYA	1966-01-05	531.119995	531.119995	531.119995	531.119995	531.119995	0.000000e+00
4	NYA	1966-01-06	532.070007	532.070007	532.070007	532.070007	532.070007	0.000000e+00
...	...	...	...	...	...	...	...	...
13943	NYA	2021-05-24	16375.000000	16508.519530	16375.000000	16464.689450	16464.689450	2.947400e+09
13944	NYA	2021-05-25	16464.689450	16525.810550	16375.150390	16390.189450	16390.189450	3.420870e+09
13945	NYA	2021-05-26	16390.189450	16466.339840	16388.320310	16451.960940	16451.960940	3.674490e+09
13946	NYA	2021-05-27	16451.960940	16546.359380	16451.960940	16531.949220	16531.949220	5.201110e+09
13947	NYA	2021-05-28	16531.949220	16588.689450	16531.949220	16555.660160	16555.660160	4.199270e+09

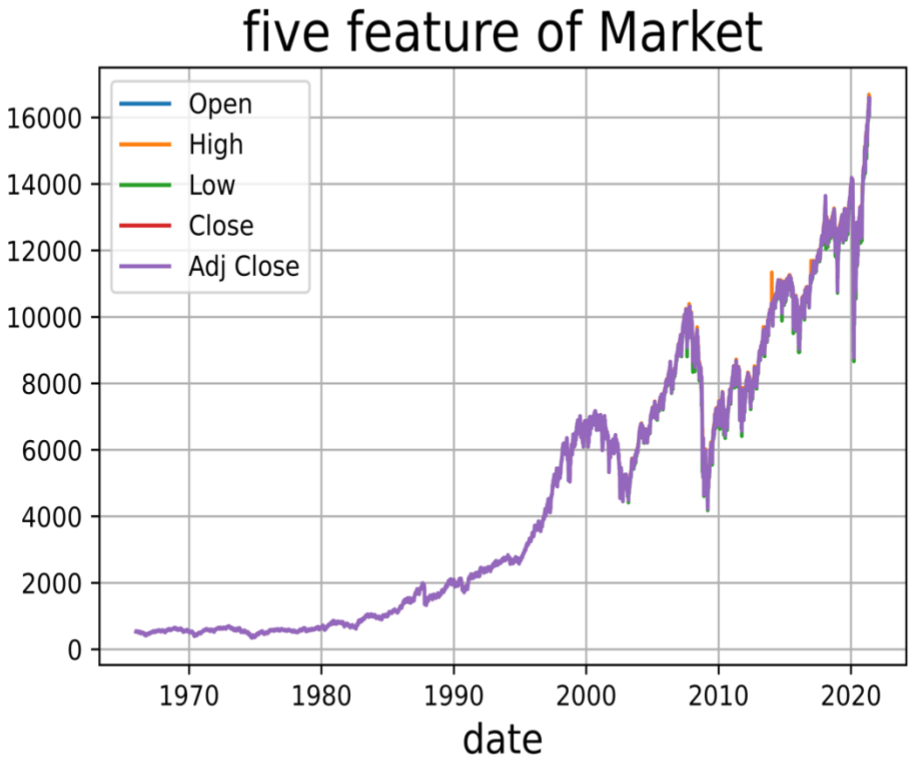
13928 rows × 8 columns

بعد از بررسی و کسب اطلاعات به توجه به دیتاست زیر در سطر count پی میبریم دیتاست به درستی پاکسازی شده است:

	Open	High	Low	Close	Adj Close	Volume
count	13928.000000	13928.000000	13928.000000	13928.000000	13928.000000	1.392800e+04
mean	4457.122309	4473.603377	4439.014504	4457.519410	4457.519410	1.217224e+09
std	4075.160214	4095.378126	4053.002787	4075.665989	4075.665989	1.834855e+09
min	347.769989	347.769989	347.769989	347.769989	347.769989	0.000000e+00
25%	656.419983	656.419983	656.419983	656.419983	656.419983	0.000000e+00
50%	2634.179931	2634.179931	2634.179931	2634.179931	2634.179931	0.000000e+00
75%	7350.584961	7385.307617	7282.672363	7348.750122	7348.750122	2.687250e+09
max	16590.429690	16685.890630	16531.949220	16590.429690	16590.429690	1.145623e+10

با رسم نمودار خطی و بررسی ۵ فیچرها بر اساس تاریخ از سال ۱۹۶۵ تا ۱۹۸۳ روند نمودار تقریبا ثابت بوده و در حدود قیمتی ۵۰۰ بوده از ۱۹۸۵تا۱۹۹۵ نمودار رشد صعودی داشته است و از تقریبا ۵۰۰ به ۳۰۰۰ رسیده است و از ۱۹۹۵ تا ۲۰۰۰ شیب تندی را تجربه کرده است و از حدود قیمتی ۲۰۰۰ به حدود ۷۰۰۰ رسیده است از سال ۲۰۰۰ تا تقریبا ۲۰۰۳ نمودار نزولی بوده است و قیمت به حدود ۴۵۰۰ رسیده است و از اخرای سال ۲۰۰۳ تا نزدیک ۲۰۰۸ دوباره نمودار رشد صعودی را داشته است و قیمت به ۱۰۰۰۰ رسیده است دوباره از نزدیک سال ۲۰۰۸ تا ۲۰۰۹ شیب نزولی را

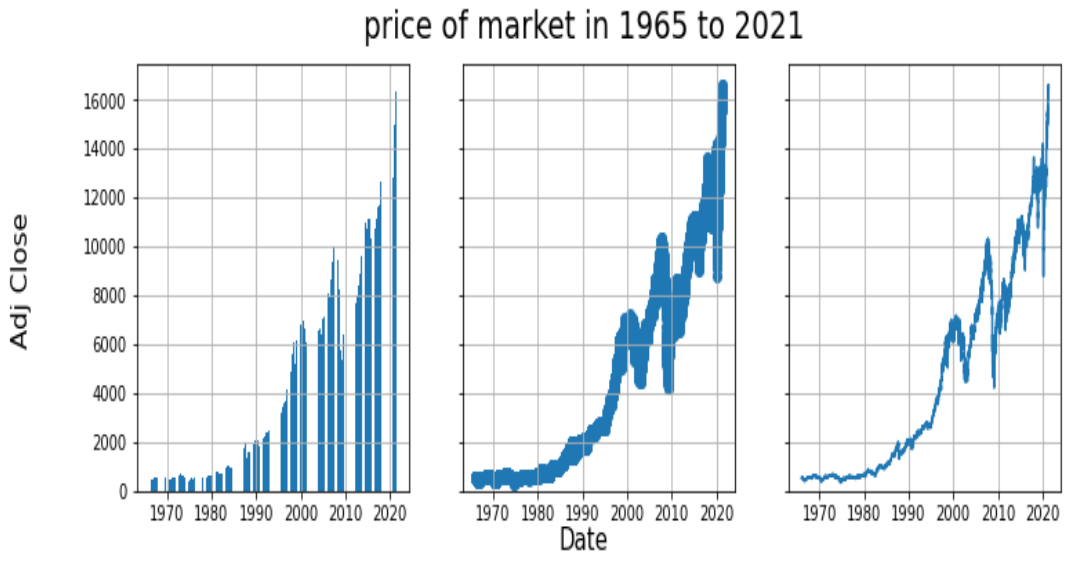
تجربه کرده است و قیمت به ۴۰۰۰ رسیده است از انتهای ۲۰۰۹ نمودار رشد صعودی خود را شروع کرده و با روند صعودی و نزولی در کل رشد قابل توجه ای را تجربه کرده و به قیمت ۱۴۰۰۰ رسیده است در ۲۰۲۰ دوباره نمودار نزولی بوده و به قیمت حدوده ۸۵۰۰ رسیده و در نهایت از ۲۰۲۰ تا ۲۰۲۱ به بالا ترین قیمت یعنی بالاتر از ۱۷۰۰۰ رسیده است!



علاوه بر نمودار قبلی ستون Adj Close از float به int تبدیل شد و سه پلات از نوع مختلف رسم شد(bar-scatter-plot) که انالیزنمودار نقطه ای و خطی در بالا توضیح داده شد اما نکته جالب در barplot به ما روند رو به رشد کلی بازار را نشان میدهد که با تمام بالا و پایین ها به طور کلی با گذشت زمان روند افزایشی بوده است.

	Index	Date	Open	High	Low	Close	Adj Close	Volume
0	NYA	1965-12-31	528.690002	528.690002	528.690002	528.690002	528	0.000000e+00
1	NYA	1966-01-03	527.210022	527.210022	527.210022	527.210022	527	0.000000e+00
2	NYA	1966-01-04	527.840027	527.840027	527.840027	527.840027	527	0.000000e+00
3	NYA	1966-01-05	531.119995	531.119995	531.119995	531.119995	531	0.000000e+00
4	NYA	1966-01-06	532.070007	532.070007	532.070007	532.070007	532	0.000000e+00
...	...	...	...	...	...	...	...	...
13943	NYA	2021-05-24	16375.000000	16508.519530	16375.000000	16464.689450	16464	2.947400e+09
13944	NYA	2021-05-25	16464.689450	16525.810550	16375.150390	16390.189450	16390	3.420870e+09
13945	NYA	2021-05-26	16390.189450	16466.339840	16388.320310	16451.960940	16451	3.674490e+09
13946	NYA	2021-05-27	16451.960940	16546.359380	16451.960940	16531.949220	16531	5.201110e+09
13947	NYA	2021-05-28	16531.949220	16588.689450	16531.949220	16555.660160	16555	4.199270e+09

13928 rows × 8 columns



البته برای انالیز داده های مالی باید از نمودار مخصوص داده مالی استفاده کرد که صعودی بودن و نزولی بودن را با کندل ها رنگی مشخص کرد که کندل سبز سعودی و قرمز نزولی است و در این نمودار میشود به درستی بازار را زیر نظر داشت و از اصلاح بازار روند کلی آن با خبر شد. در اینجا تا سال ۲۰۰۴ روند سعودی بوده دقیقا شبیه عرضه اولیه از ۲۰۰۴ تقریبا با کندل های صعودی و نزولی در کل رشد داشته تقریبا از ۲۰۰۸ ریزش را تجربه کرده است و از ۱۰۰۰۰ به قیمت ۴۰۰۰ رسیده و همینطور اصلاحاتی را در روند خود تجربه کرده است