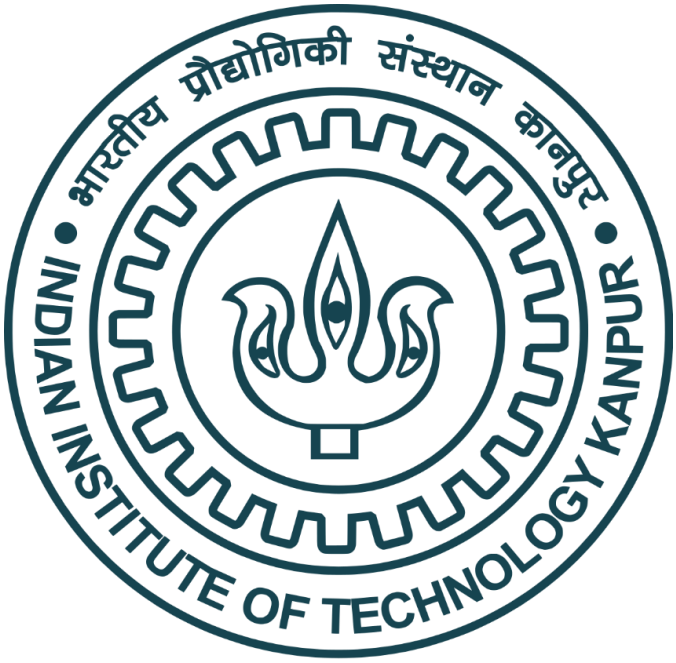
Department Of Mathematics and Statistics

MTH-517A Time Series Analysis

Semester – III



Course Project

Analyzing And forecasting using the Amazon Stock Market Data

Submitted By –

Gagan Deep 211303

Gaurav 211304

Prithwijit Ghosh 211349

Sarvesh Singh Kushwaha 211455

Supervised By – Dr. Amit Mitra

Introduction

A stock market, equity market, or share market is the aggregation of buyers and sellers of stocks (also called shares), which represent ownership claims on businesses; these may include securities listed on a public stock exchange, as well as stock that is only traded privately, such as shares of private companies which are sold to investors through equity confounding platforms. The stock market is very volatile and unpredictable, and a small geographical or socio-economical change can impact the share trends of stocks in the stock market, recently we have seen how COVID-19 has impacted the stock market. The randomness and forever fluctuating market make investing a risky business. In order to get ahead of the market, statisticians have always tried to analyze and forecast the stock price that reflects all known and unknown information in the public domain.

The goal of this project is to analyze Amazon's stock price data for the time period of 1997-2020. We have analyzed Amazon's opening price movement and fitted a suitable model for forecasting.

INDEX

Dataset information

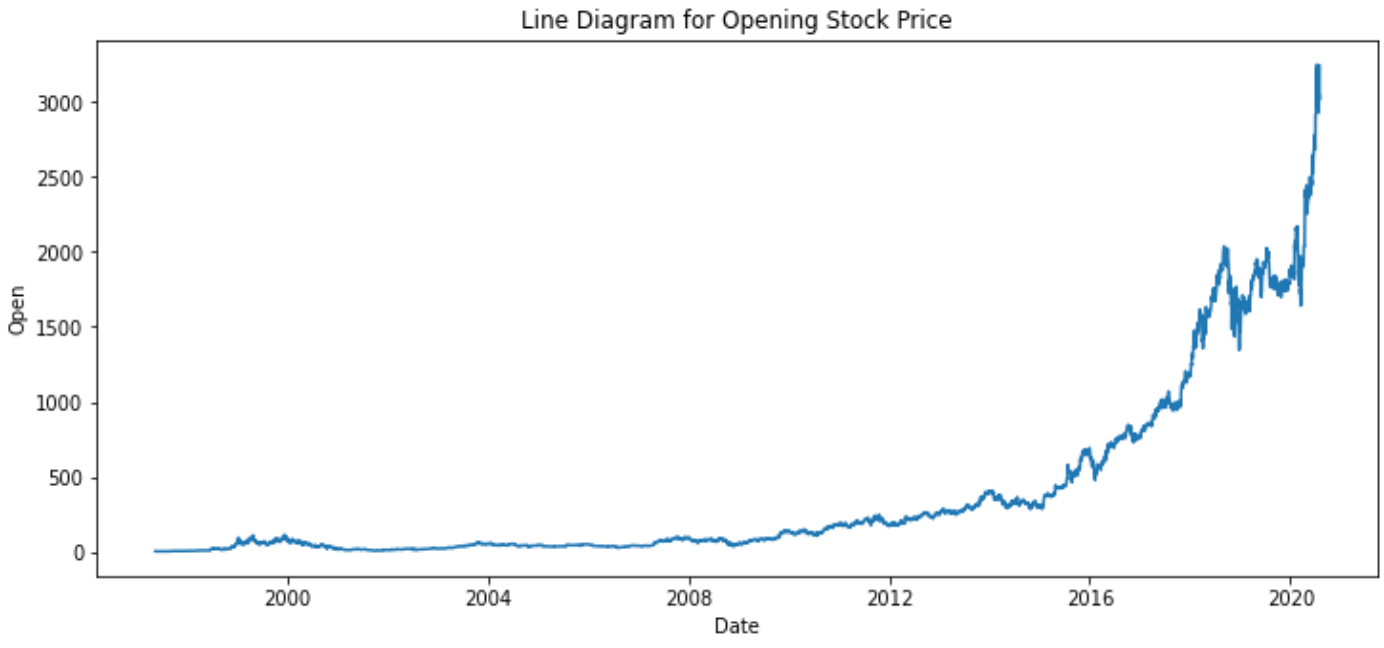
Amazon Stock Prices from 1997 to 2020 data has 7 variables we described all variable:-

Variable	**Description**
Date	in the format: yy-mm-dd.
Open	the price of the stock at market open.
High	The highest price reached in the day.
Low	The lowest price reached in the day.
Close	The stock closes at the end of the Market hours
Adj Close	This is the closing price after adjustments for all applicable splits and dividend distributions.
Volume	The number of shares traded.

Plotting Data Set

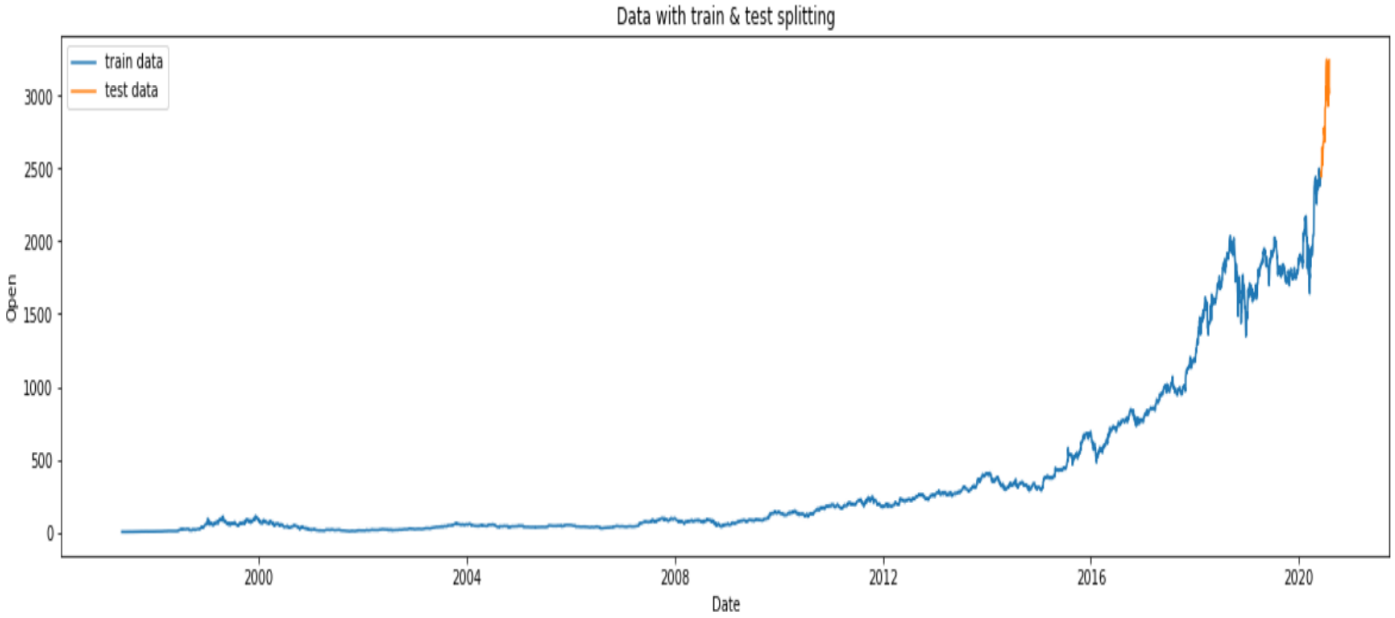
We plotted the graph between the year 1997 to 2020 and the opening stock price we observed that from 1997 to 2012 opening stock prices has relatively constant then we have seen from the 2012 to 2019 graph exponentially increase, and from the year 2019 opening stock price linearly increases.

As we have seen in the graph from 2019 to 2020 opening stock prices have been falling. Due to some economic causes.



Train Test Splitting

For forecasting, we need to split the data set into train and test data, first 5800 data points we have used as training data, and 42 data points we used as test data.



We plotted the graph of the training and test data set graph blue line shows the training data set which we use for fitting the model and the yellow line shows the test data set which is used for the prediction.

Testing for the existence of trend (Relative Ordering Test)

This is a non-parametric test procedure used for testing the existence of a trend component.
Null hypothesis against Alternative hypothesis

H_0: No Trend

against H_A: Trend is present

Let the time series be denoting by{ Y_1, Y_2 ……, Y_N} (at 'n' time points)

Define:

$$q_{ij} = \begin{cases} 1, & \text{if } y_i > y_j \text{ when } i < j \\ 0, & o/w \end{cases}$$

$$Q = \sum_{\substack{i \\ i<j}} \sum_{j} q_{ij}$$

Note that **Q** counts the number of decreasing points in the time series and is also the number of discordances.

If there is no trend (increasing or decreasing) in the time series.

$$P(q_{ij} = 0) = P(q_{ij} = 1) = \frac{1}{2}$$

Under no trend (i.e. under H_0)

$$E(Q) = \sum \sum_{i<j} E(q_{ij}) = \frac{n(n-1)}{4}$$

if observed Q<< E(Q) then it would be an indication of rising trend and if observed Q >> E(Q) then it would be an indication of a falling trend.

If observed Q does not differ "significantly" from E(Q) (under H_0) then it would indicate no trend.

Q is related with Kendall's T (tau) , the rank correlation coefficient, through the relationship.

$$T = 1 - \frac{4\mathbb{Q}}{n(n-1)}$$

using the standard results of Kendall's T, we have that, under the null hypothesis of no trend.

$$E(T) = 0 \quad \text{and} \quad V(T) = \frac{2[2n+5]}{9n(n-1)}$$

Asymptotic test for H_0: no trend is based on the statistic

$$Z = \frac{T - E(T)}{\sqrt{v(T)}} \sim N(0,1)$$

We would reject the null hypothesis of no trend at level of significance α if observed

$$|Z| > T^{\alpha}/_{2}$$

($T\alpha/\mathbf{2}$ is the $\alpha/2$ the upper cut-off point of a standard normal distribution i.e.

$$P(Z > T\alpha\ /\ 2\ ) = {}^{\alpha}/_{2} \qquad \text{Where } Z \sim N(0,1)$$

We applied this method at 95% level of significance then we got Z = 1.65

After applying the hypothesis, observed value is greater than the expected value then we get: -

The Null Hypothesis is rejected. So, trend is present.

Trend Elimination by Differencing Method

First, we have taken the logarithm of the data then we apply Trend Elimination by differencing method.

This is a method of trend elimination without estimating the trend component.

In order to remove trend, we applied the difference operator of lag 2 on our data. In most of the cases, the stock price are assumed to dependent in the immediate previous value.

Define

Lag operator : B : For first lag operator

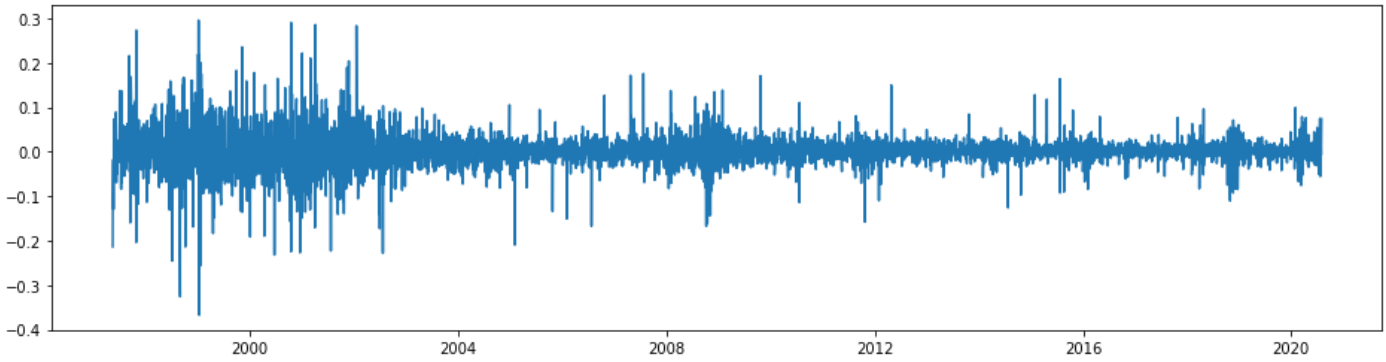
$$B\ Y_t = Y_{t-1},$$

First deference operator:

$$\nabla\ Y_t = Y_t - Y_{t-1}$$

After applying 2nd order difference operator the data gets detrended. Which is shown below

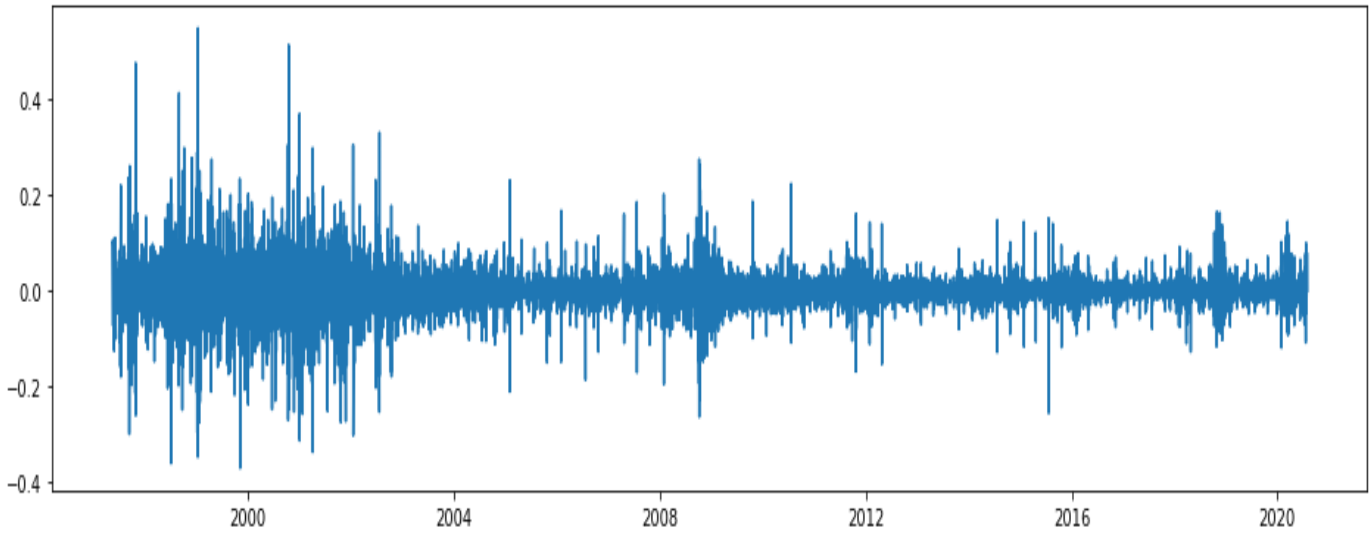
Date	Open	log	Delta1	Delta2
1997-05-15 00:00:00+00:00	2.437500	0.890973	-0.213574	0.101726
1997-05-16 00:00:00+00:00	1.968750	0.677399	-0.111848	0.093937
1997-05-19 00:00:00+00:00	1.760417	0.565551	-0.017911	-0.037831
1997-05-20 00:00:00+00:00	1.729167	0.547640	-0.055742	-0.073250
1997-05-21 00:00:00+00:00	1.635417	0.491898	-0.128992	0.107013
...	...	...	...	...
2020-07-27 00:00:00+00:00	3062.000000	8.026824	-0.002528	-0.005124
2020-07-28 00:00:00+00:00	3054.270020	8.024296	-0.007651	0.002030
2020-07-29 00:00:00+00:00	3030.989990	8.016645	-0.005621	0.079160
2020-07-30 00:00:00+00:00	3014.000000	8.011023	0.073539	0.000000
2020-07-31 00:00:00+00:00	3244.000000	8.084562	0.073539	0.000000



After applying First difference order operator :

For checking the presence of the trend we have applied the Relative order test.
We obtained Z>1.65 [<T0.95(=2.28)], hence we can reject the null hypothesis, so we conclude that trend is present in the data.

After applying Second difference order operator :



For checking the presence of the trend we have applied the Relative order test.
We obtained Z>1.65[>$T_{0.95}$(=1.64)],so we can accept the null hypothesis of no trend,
Hence we conclude that trend is not present in the data.

The following figure shows the series after applying Difference operator

The figure shows both trended and detrended data

From the graph we can conclude that, the trend has been removed after applying the 2nd-order difference operator.

Dealing With Seasonality

We converted our data in the monthly format in order to test for seasonality-

Month	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	...
1	0.000000	0.064447	0.015280	-0.078502	0.113312	-0.093003	-0.019607	-0.015272	0.068038	0.002721	...
2	0.000000	-0.081355	0.041225	-0.034798	0.056732	0.152177	-0.029275	0.026668	-0.004541	-0.056205	...
3	0.000000	0.087645	-0.105393	0.008647	-0.203448	-0.038293	0.042314	0.042796	-0.030599	0.026371	...
4	0.000000	-0.040414	0.004700	0.076695	0.258376	-0.053521	-0.034827	-0.012100	0.038607	0.009090	...
5	0.102507	-0.032794	0.015898	0.018225	-0.065382	0.004110	-0.041076	0.005388	0.002589	0.023772	...
6	-0.053928	0.136362	0.034310	-0.055402	0.001630	-0.142206	0.085378	-0.051714	-0.014127	-0.017989	...
7	0.067008	-0.113214	0.019714	-0.004428	-0.009462	0.139325	-0.068862	0.026509	-0.020443	0.008468	...
8	-0.051230	0.125619	-0.110316	-0.040485	-0.014261	0.010520	0.060681	-0.012324	0.003394	0.011769	...
9	-0.039805	-0.151281	0.113784	0.023766	0.058436	0.040719	-0.020168	0.012058	0.026180	-0.055154	...
10	0.048544	0.032730	0.024051	0.071987	-0.062998	-0.024055	0.035316	0.022325	0.003594	0.008853	...
11	0.033985	0.084648	-0.066323	-0.043312	-0.014671	-0.014872	-0.037940	-0.042626	-0.018570	0.014401	...
12	-0.018937	-0.120890	0.110303	-0.164161	0.040183	0.033416	-0.011898	-0.041808	0.009193	0.017413	...

12 rows × 25 columns

Testing for seasonality using Friedman's test-

STEP-1 – Trend component which was present has been removed

Step 2- Ranking the values obtained from step 1 within each year from smallest to largest. We get the following result-

Month	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	...	2012	2013	2014	2015	2016	2017	2018	2019	2020	Mi
1	6.5	8.0	5.0	5.0	11.0	2.0	4.0	5.0	12.0	6.0	...	11.0	1.0	3.0	12.0	6.0	8.0	7.0	1.0	8.0	0.0
2	6.5	4.0	11.0	4.0	10.0	12.0	5.0	11.0	9.0	2.0	...	1.0	12.0	11.0	2.0	12.0	4.0	1.0	10.0	10.0	0.0
3	6.5	9.0	1.0	9.0	1.0	3.0	10.0	12.0	1.0	11.0	...	7.0	3.0	8.0	3.0	1.0	7.0	8.0	7.0	1.0	0.0
4	6.5	5.0	3.0	11.0	12.0	4.0	6.0	4.0	11.0	7.0	...	4.0	5.0	6.0	9.0	8.0	11.0	10.0	5.0	2.0	0.0
5	12.0	6.0	8.0	10.0	2.0	7.0	2.0	9.0	8.0	10.0	...	6.0	9.0	1.0	7.0	3.0	1.0	5.0	4.0	11.0	0.0
6	2.0	12.0	9.0	2.0	7.0	1.0	12.0	1.0	2.0	3.0	...	8.0	10.0	12.0	6.0	10.0	3.0	9.0	11.0	12.0	0.0
7	9.0	1.0	7.0	6.0	4.0	11.0	1.0	8.0	4.0	4.0	...	2.0	6.0	2.0	4.0	4.0	12.0	3.0	6.0	9.0	0.0
8	4.0	10.0	2.0	7.0	5.0	8.0	11.0	6.0	3.0	12.0	...	10.0	8.0	10.0	11.0	5.0	2.0	6.0	12.0	5.0	0.0
9	1.0	2.0	12.0	8.0	9.0	10.0	8.0	7.0	10.0	1.0	...	5.0	7.0	5.0	5.0	7.0	6.0	4.0	3.0	5.0	0.0
10	11.0	7.0	6.0	12.0	3.0	5.0	9.0	10.0	6.0	8.0	...	9.0	2.0	7.0	8.0	2.0	5.0	11.0	8.0	5.0	0.0
11	10.0	11.0	4.0	3.0	6.0	6.0	3.0	2.0	5.0	5.0	...	3.0	4.0	4.0	10.0	9.0	9.0	2.0	2.0	5.0	0.0
12	3.0	3.0	10.0	1.0	8.0	9.0	7.0	3.0	7.0	9.0	...	12.0	11.0	9.0	1.0	11.0	10.0	12.0	9.0	5.0	0.0

12 rows × 25 columns

Step 3- Computing the asymptotic test static

$$X = 12 * \sum_{i=1}^{12} \frac{\left(M_i - \frac{c(r+1)}{2}\right)^2}{cr(r+1)} i$$

We can say that the test statistics $X \sim \chi^2_{r-1}$

So, the asymptotic test would reject null hypothesis of no seasonality at α level of significance if

Observed(X) > $\chi^2_{r-1}(\alpha)$

Where $\chi^2_{r-1}(\alpha)$ is the upper α cutoff point of a centsquare distributionχ^2 distribution on r-1 degrees of freedom,

i.e. P(χ^2 > $\chi^2_{r-1}(\alpha)$) = α

So, our test statistics value and the value of $\chi^2_{r-1}(\alpha)$ (r=12) is-

```
Obesrved X / Value of Test Statistics= 11.717948717948715
Value of central chi-square(r-1) at α=5% is = 19.67513757268249
```

Hence, found out that

```
The Null Hypothesis is accepted....so seasonality is not presented
```

Dealing With Randomness

We will use the turning point test which is a non-parametric test for testing the randomness of a time series.

We identify a point as a turning point if it is greater than or less than its 2 neighboring values.

Number of points satisfying the condition for turning point = 4310

Defining our test statistics Z:

Z = $\frac{P-E(P)}{\sqrt{V(P)}} \sim N(0,1)$ { Under the Null Hypothesis that Series is purely random}

We would reject H_o at α level of significance if observed |Z| > $\tau_{\alpha/2}$s

The value of our test statistics comes out to be 12.933349159623724 which is much higher than 1.959963984540054

»H_o is rejected.

Hence, our data is not random.

Test For Stationarity – ADF Test

Augmented Dickey-Fuller Test (ADF) tests for the stationarity of the data.
It tests for the null hypothesis that a unit root is present in a time series i.e. the series is not stationary, against that the series is Stationary
If we fail to reject the null hypothesis it means the time series has a unit root i.e. the time series is not stationary
If the null hypothesis is rejected, it means the time series does not have a unit root i.e. the series is stationary
For testing, we compare the p-value obtained by the Augmented Dickey-Fuller Test with the α level of significance.
If the p-value is smaller than the α L.O.S, the null hypothesis will be rejected Hence the series is stationary
And if the p-value is bigger than α L.O.S, then we failed to reject the null hypothesis Hence the series is not stationary.

```
p-value -  0.0
alpha -  0.05
p-value -  0.0  < aplha = 0.05
So we reject the null hypothesis
```

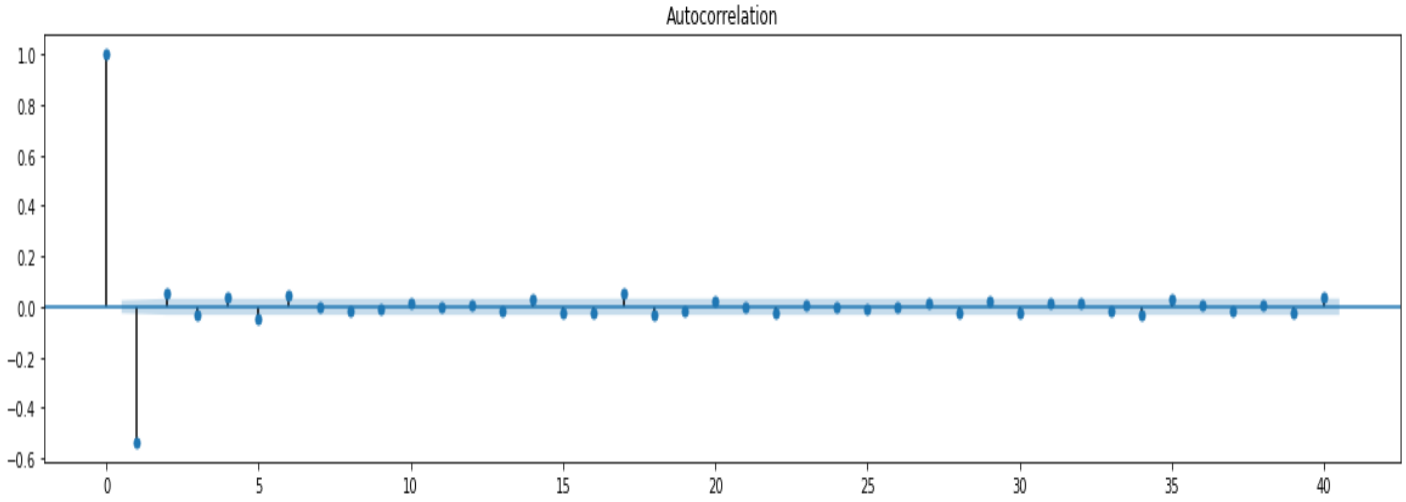
We get the p-value = 0.0 < 0.05 = alpha So we reject the null hypothesis, i.e. the time series does not have a unit root.
Hence the time series is stationary.

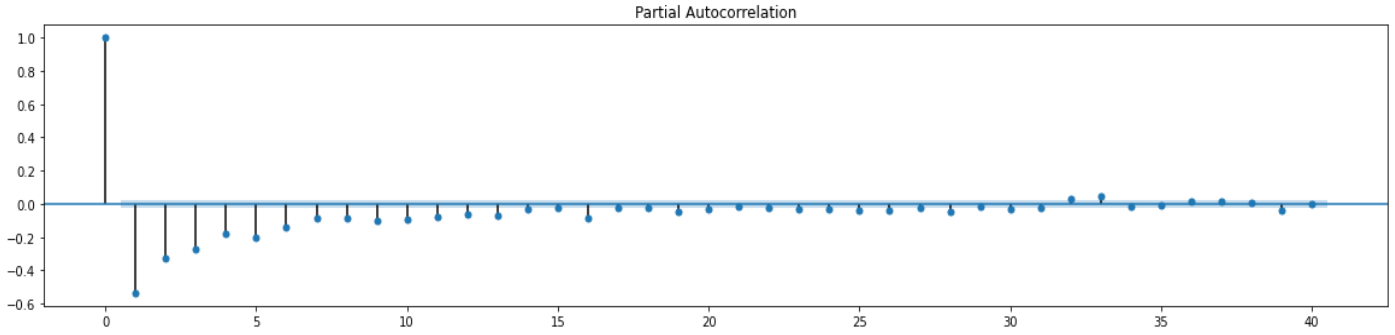
ACF and PACF plots for deciding order of the model :

Autocorrelation and partial autocorrelation plots are used to graphically summarize the strength of a relationship between an observation of a time series and observations at prior time steps. Plots of autocorrelation function(ACF) and partial autocorrelation function(PACF) give us different viewpoint of time series.

A partial autocorrelation plot is a summary of the relationship between an observation in time series with observation at prior time steps with the relationships of intervening observation removed i.e. PACF only describes the direct relationship between an observation and its lag. This would suggest that there would be no correlation for lag values beyond k. while ACF describes the autocorrelation between an observation and another observation at a prior time step that includes direct and indirect dependence information,

We use the ACF and PACF plots to determine the order of AR and MA. For both the plots, we take that value of lag, After which value of the ACF and PACF are not significantly different from the ACF plot, we get the order MA(q) i.e. the value of ‘q’ and from PACF plot, we get the order of AR(p), i.e. the value of ‘p’.





The graphs we can see that, ACF plot cut-off after at the lag order of 1 and PACF plot Tail-off same lag order. Then we can clearly say that it is a MA(1) process
So we will continue with the model fitting of MA(1).

Forecasting

The primary objective of building and fitting a model in a time series is to predict the future time points and also to check the precision on these data points.

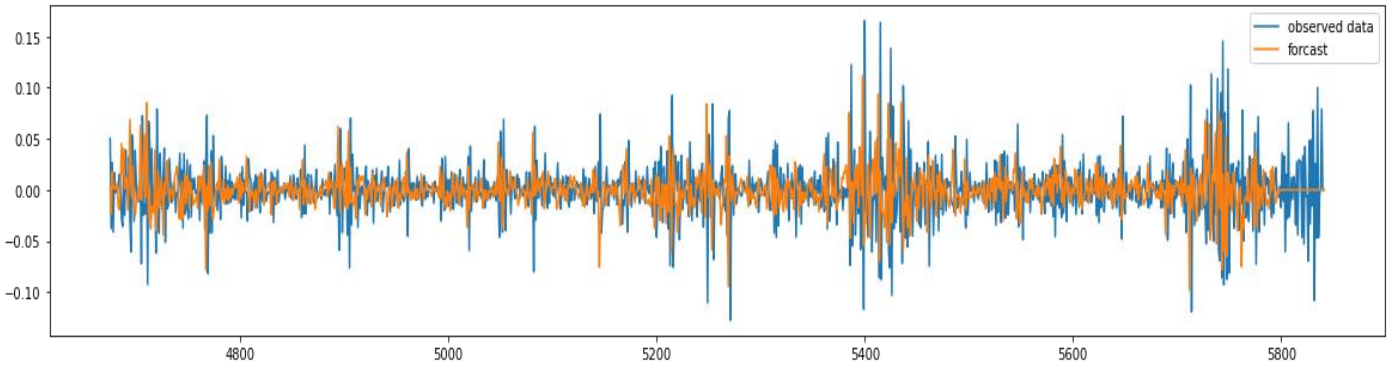
FITTING THE MA(1) MODEL

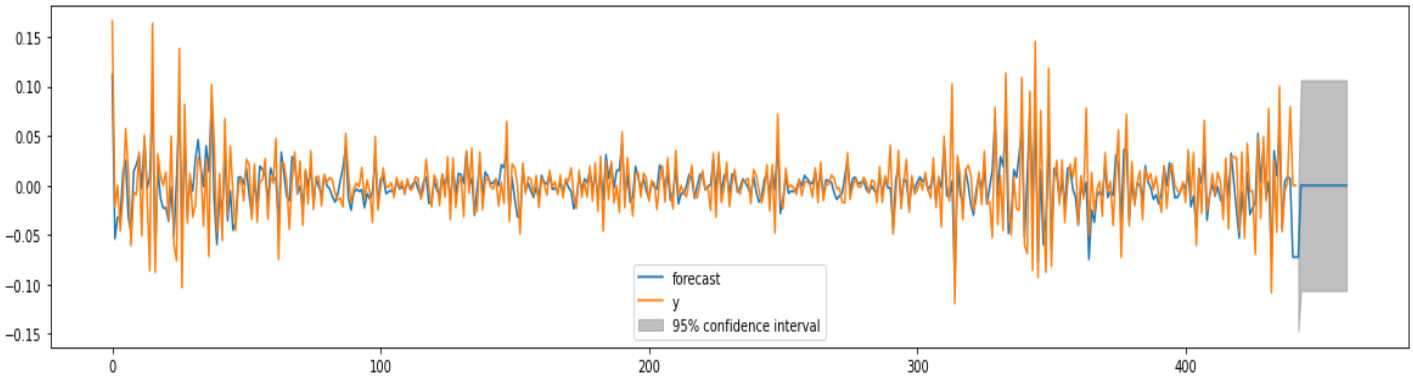
As we have seen earlier that MA(1) is best suited for our data, we estimate the parameters for this model, where the model is –

$$x_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}$$

Where Xt is our original data and $\boldsymbol{\varepsilon_t}$ is the white noise for the t^{th} time point.

So the corresponding estimastes are 0.99 and 0.





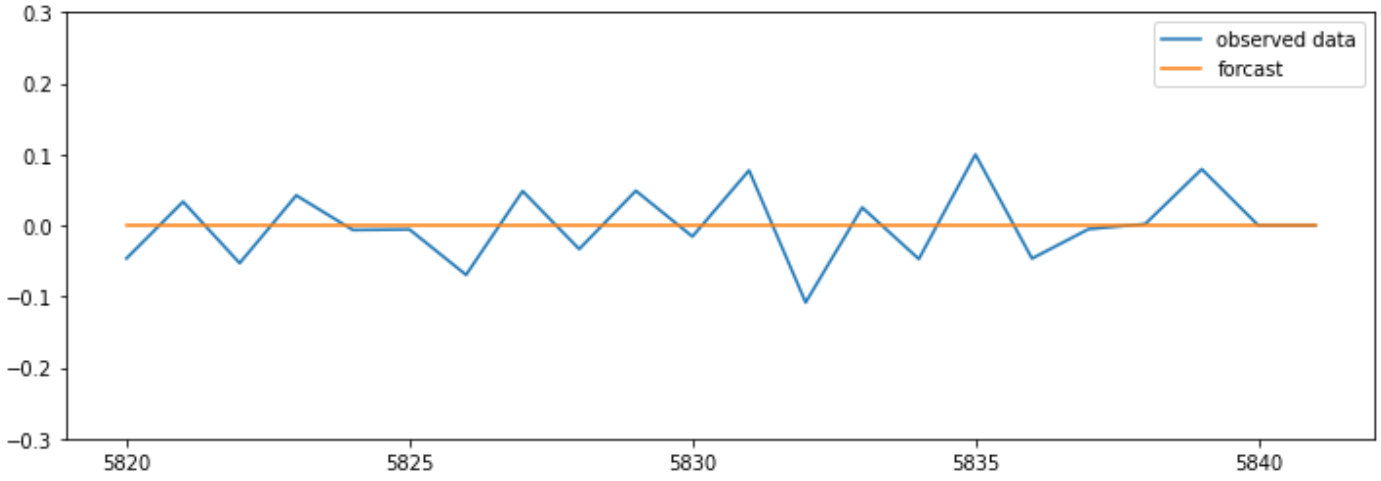
Forecasting in stationary time series: Best Linear Predictor (BLP)

$\{x_t\}$ - Covariance stationary time serins with mean μ and ACVF $\{\gamma_h\}$

Given information upto time $n, \{x_1, \ldots, x_n\}$ problem is to predict x_{n+h} for some $h > 0$.

BLP approach:
Find the linear combination of $x_n, \dots, x_1$ that provides the "best" forecast of x_{n+h} "best" : w.r.t. minimum mean square prediction error

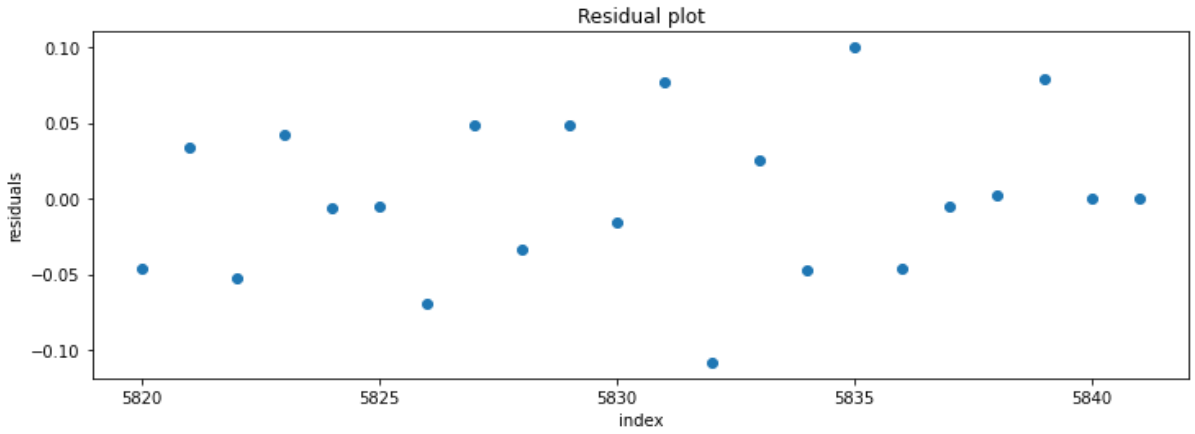
Def": Best Linear Predictor (BLP)
BLP of x_{n+h} in terms of $(x_n, x_{n-1}, \dots, x_1)$ denoted by $P_{(x_n,\dots,x_1)}x_{n+h} = p_n x_{n+h}$ is the linear f^n: $a_0^* + a_1^* x_n + \cdots + a_n^* x_1$ if $E(x_{n+h} - p_{(x_n,\dots,x_1)}x_{n+h})^2$ minimum among all such linear functions.



Residual Analysis

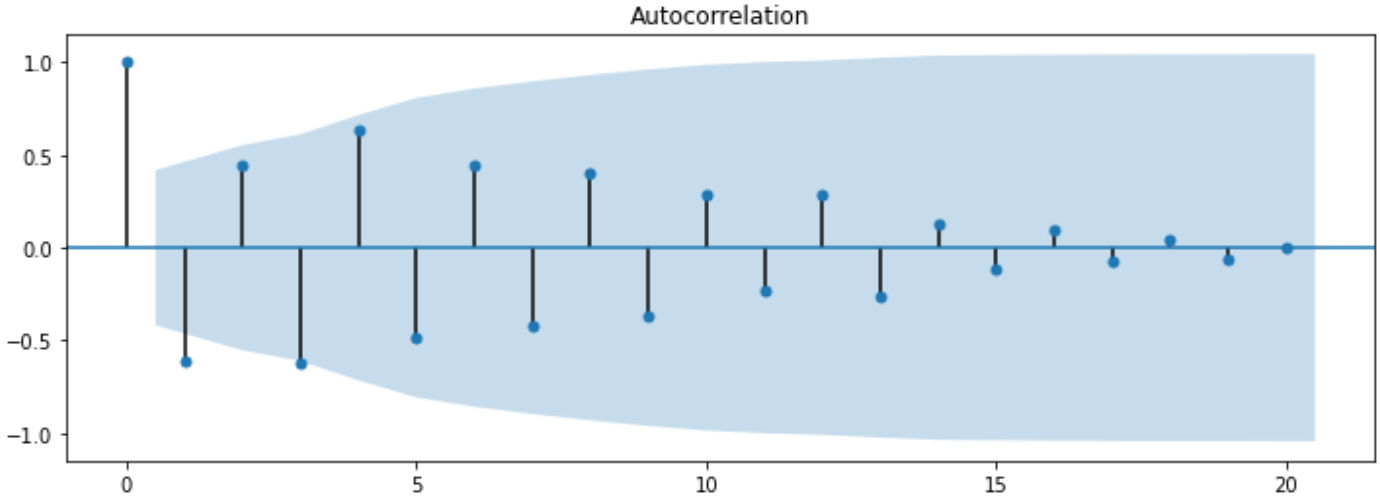
The above BLP model doesn't look very good for fitting the data. Hence, we will study the residuals for this fitted and observed model.

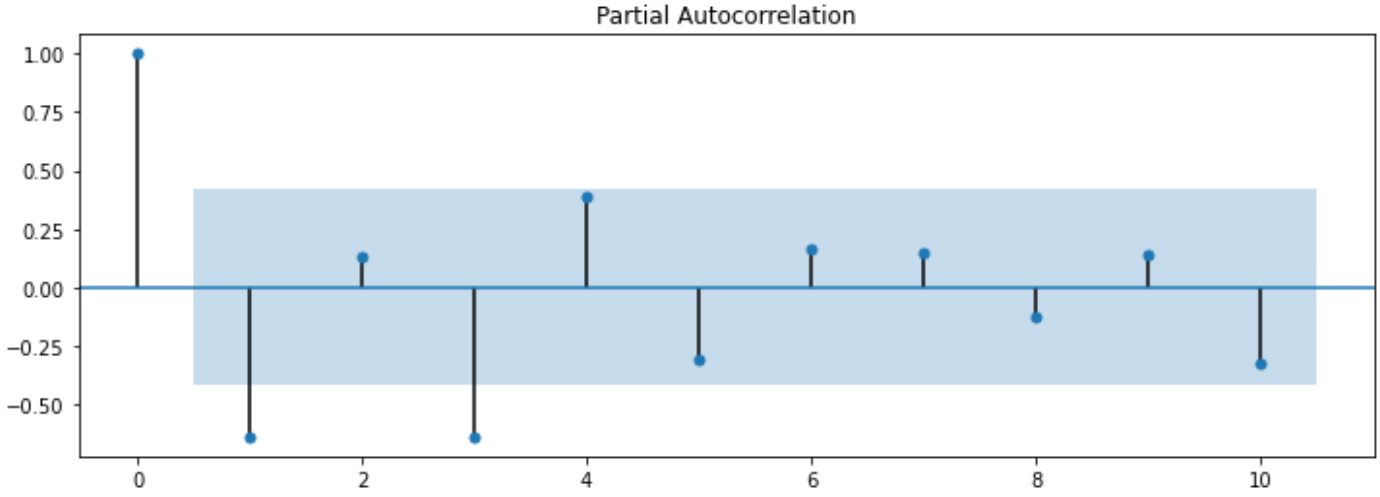
The most obvious plot for checking the randomness is scatter plot corresponding to their indices.



We see that there is no regular pattern in the scattar plot so, the residuals seems to be random.

Now, we have to check by the autocorrelation plot for the interrelated dependency.





From the ACF and PACF we observe that only a little amount of correlation is present in our data.

Hence, we can conclude that our residuals are random.

CONCLUSION

We choose tha opening stock price as our time series variable, then we eliminated all the necessary deterministic components from the data. Finally, we fitted the population stationary model (MA(1)) by observing the acf and pacf and we forcasted the test points fitting the blp model. But, our forcasted data doesn't look good. But, the corresponding residual analysis says that the residuals are random.