Predicting the Future: A Time Series Analysis of Nifty 50 Stock Prices

Project Submitted by :

Abhinav Kumar (221252)

Akash Rawat (221259)

Anupam Anand (221277)

Deevanshu Saini (221311)

Shubham Kr. Rajak (221423)

INTRODUCTION

The project is to forecast the stock price of Nifty50 will increase or decrease based on the historical idea.

Below is the google drive link:-

https://drive.google.com/drive/folders/1ypZBDn5cL_E2Du0m1mMKVqJ56q5-y97s

In this project we intent to :

- Impute the missing values in the data using interpolation method.
- Detect the presence of Trend and Seasonality in the data.
- Remove the systematic components in order to obtain the residuals.
- Check the Stationarity of the residuals using Dicky-Fuller test.
- Fit an ARIMA model to the data.
- Forecast the future outcomes using the model.

Data Overview

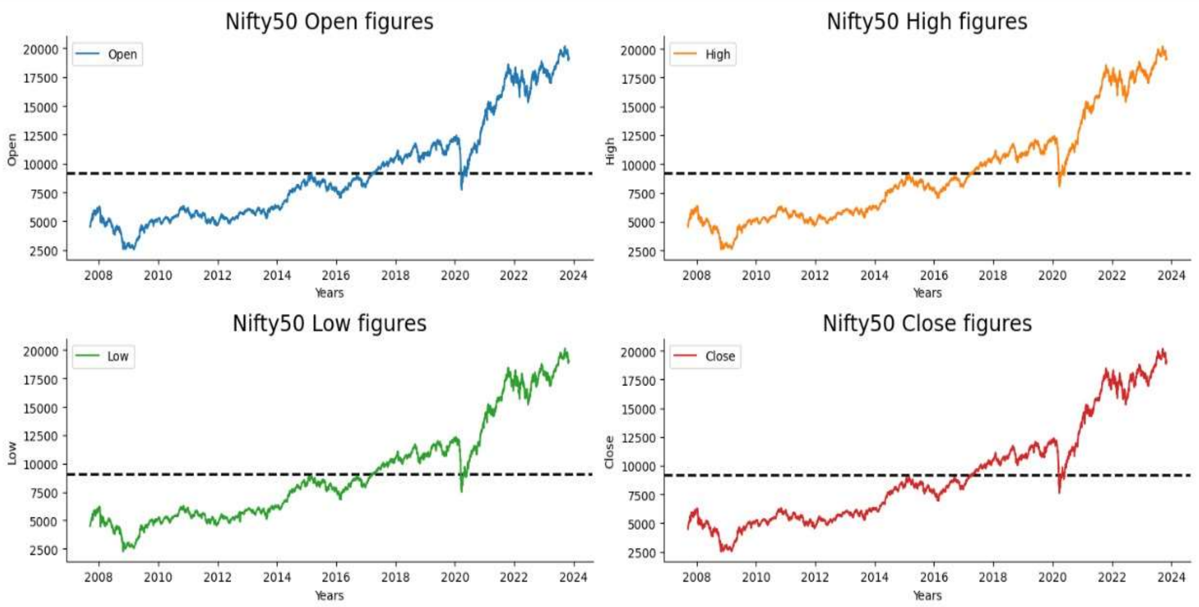
The dataset has the following features.

- **Date** - Each trading day
- **Open** - Open price of stock
- **High** - High price of stock in the particular day
- **Low** - Low price of the stock in the particular day
- **Close** - Close price of the stock at end of the day
- **Volume** - Volume traded in the entire day

Missing Values

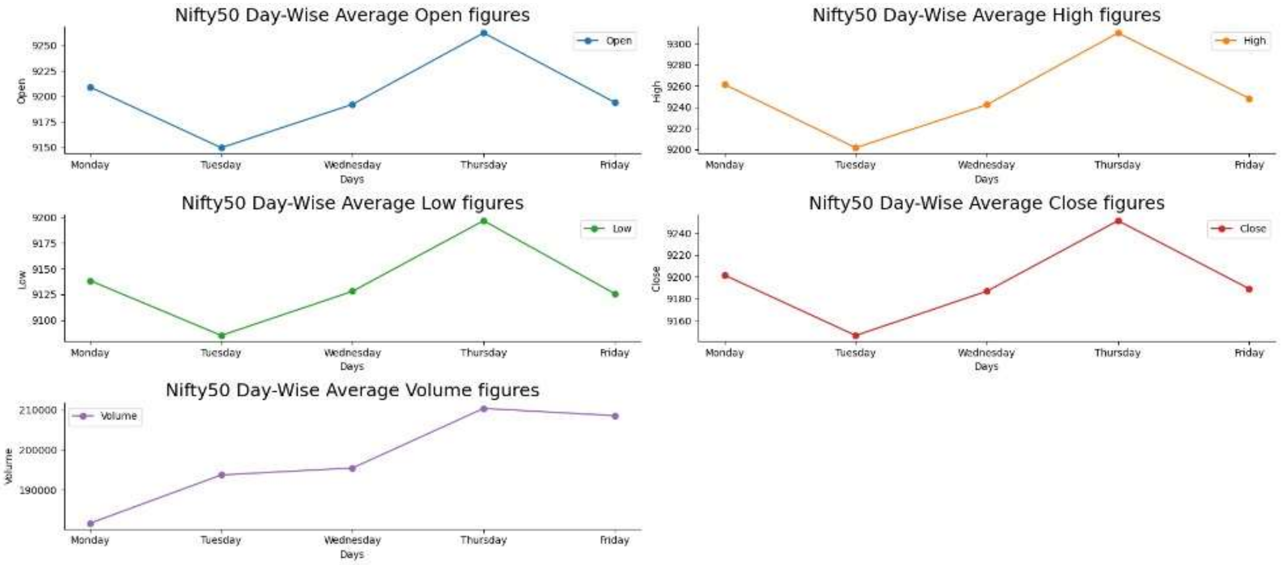
Yes, the Volume component has no previous available data for some time frame. Hence we imputed them by filling the values 0.

View of Raw Data



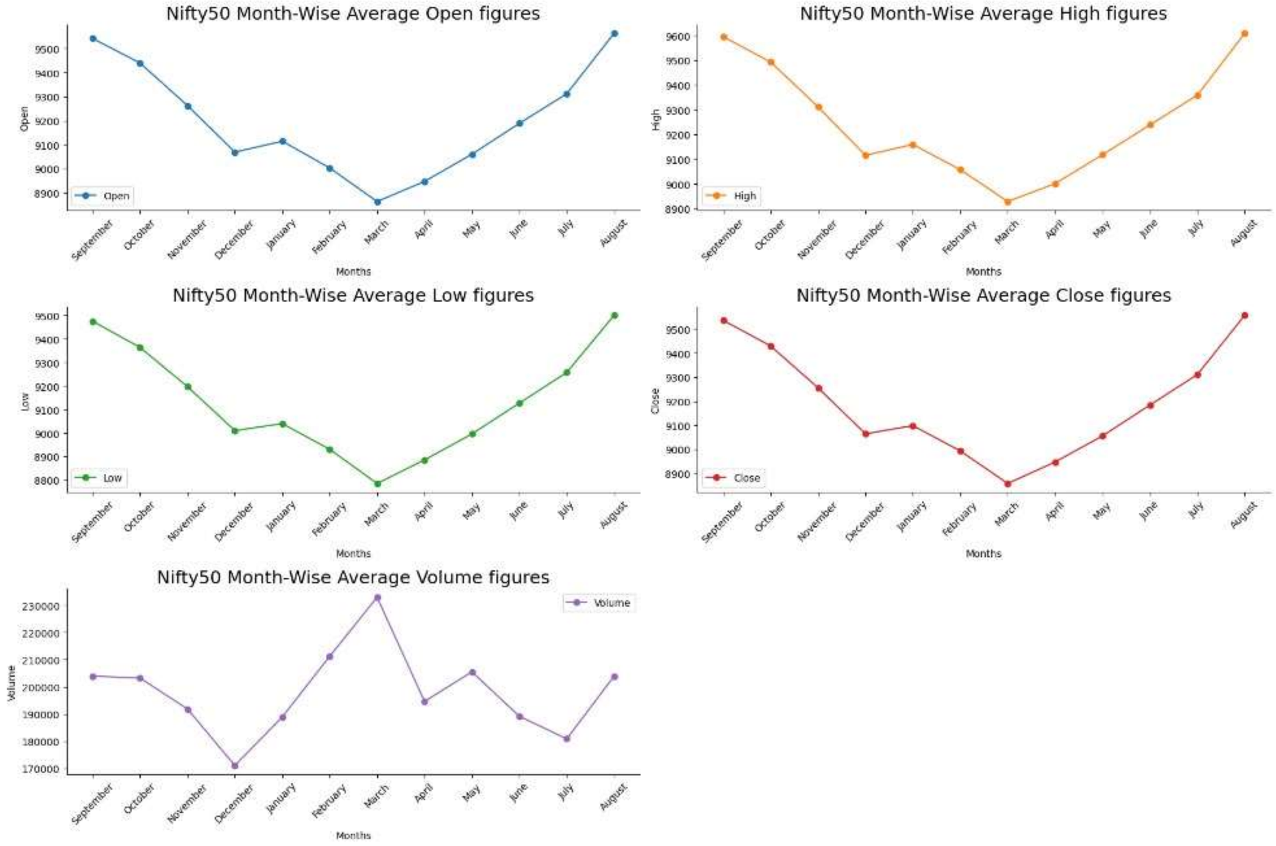
Exploratory Data Analysis

Plot for day wise :-



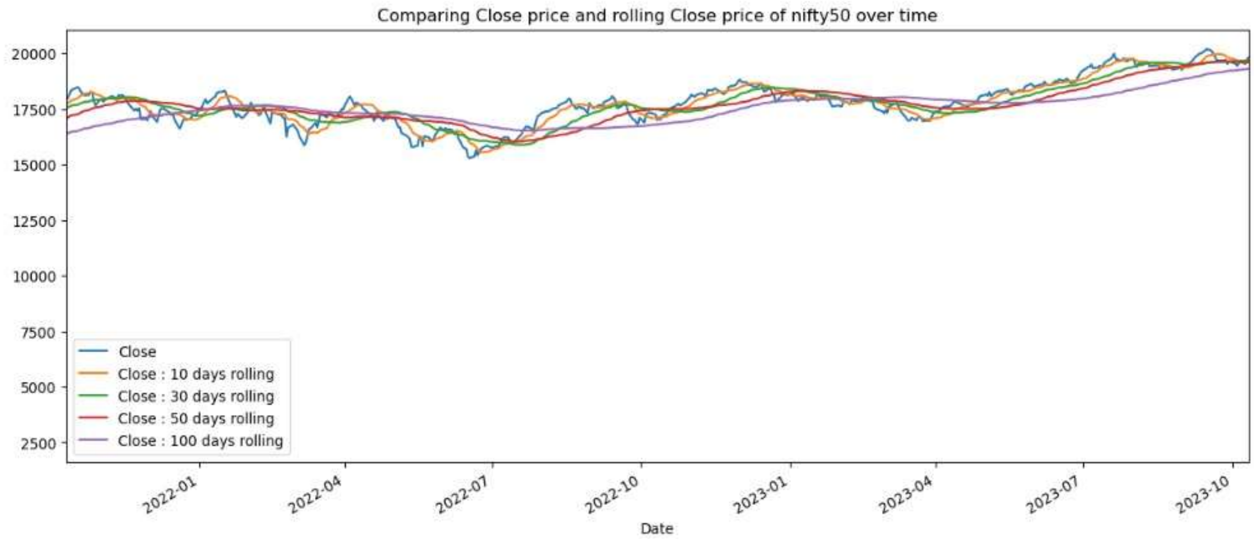
Observation: On Thurday the average prices get high where as on Tuesday the average price get low.

Plot for month wise :-



We can observe here that the average price significantly decreased in March, while the volume increased.

Plot for simple moving average :-



Check for randomness in Data

At first we need to check the randomness of the data.

So, the testing of hypothesis problem will be:

H_0: Data is purely random v/s H_1: Data is not random.

Now, to test the Null hypothesis we have performed the *Turning point test* (In statistical hypothesis testing, a turning point test is a statistical test of the independence of a series of random variables.)

Define, T: total number of turning points
The test statistic for testing the hypothesis will be given by: $Z= (T-E(T))/\sqrt{Var(T)}$,

where $E(T)=2/3*(n-2)$ and $V(T)=(16n-29)/90$ [under H0]

We reject the null if the Z (observed)> $Z_{0.025}$ (upper 0.025 point of Normal (0,1))

Now, the observed value of the test statistic is |Z| = ****** > 1.96

Variables in Dataset	Turning point test statistic value**	Upper α-points of Normal distribution
Open	-25	1.96
Close	-29.84	1.96
Low	-36	1.96
High	-35	1.96

All the features Open, Close, Low, High have irregular components.

Checking for the presence of trend

A. Testing the presence of trend:

First, we check if there is any trend component present in the data or not, where trend is referred as the long increase or decrease in the data. We check for the trend using the relative ordering test.

To test, H_0: no trend v/s H_1: H_0 not true

To calculate the test statistic we have to find out the number of discordant pairs and then Q (the total number of discordant pairs in the given time series data), where E(Q)=n*(n-1)/4 (under H_0) E(Q)= 3919410
Our observed value of Q is shown in table.

Components	Q	E(Q)
Open	586881	3919410
Close	586197	3919410
Low	584264	3919410
High	588506	3919410

So, clearly in our case Q < E(Q), providing an impression of a **rising trend** component incorporated in the data.

Also, Q is directly related to the Kendall's rank correlation coefficient (τ),

τ = 1-(4*Q/ (n*(n-1)), where E(τ)=0 and V(τ)= 2*(2n+5)/9n(n-1)

Now, we can define the test statistics as: Z= (τ -E(τ))/√Var(τ) ~ N(0,1) (under H0)

Based on the asymptotic test the value of test statistic |Z| = ** > 1.96 (upper 0.025 point of Normal(0,1).

Table for test statistic value and upper α-points of Normal distribution:

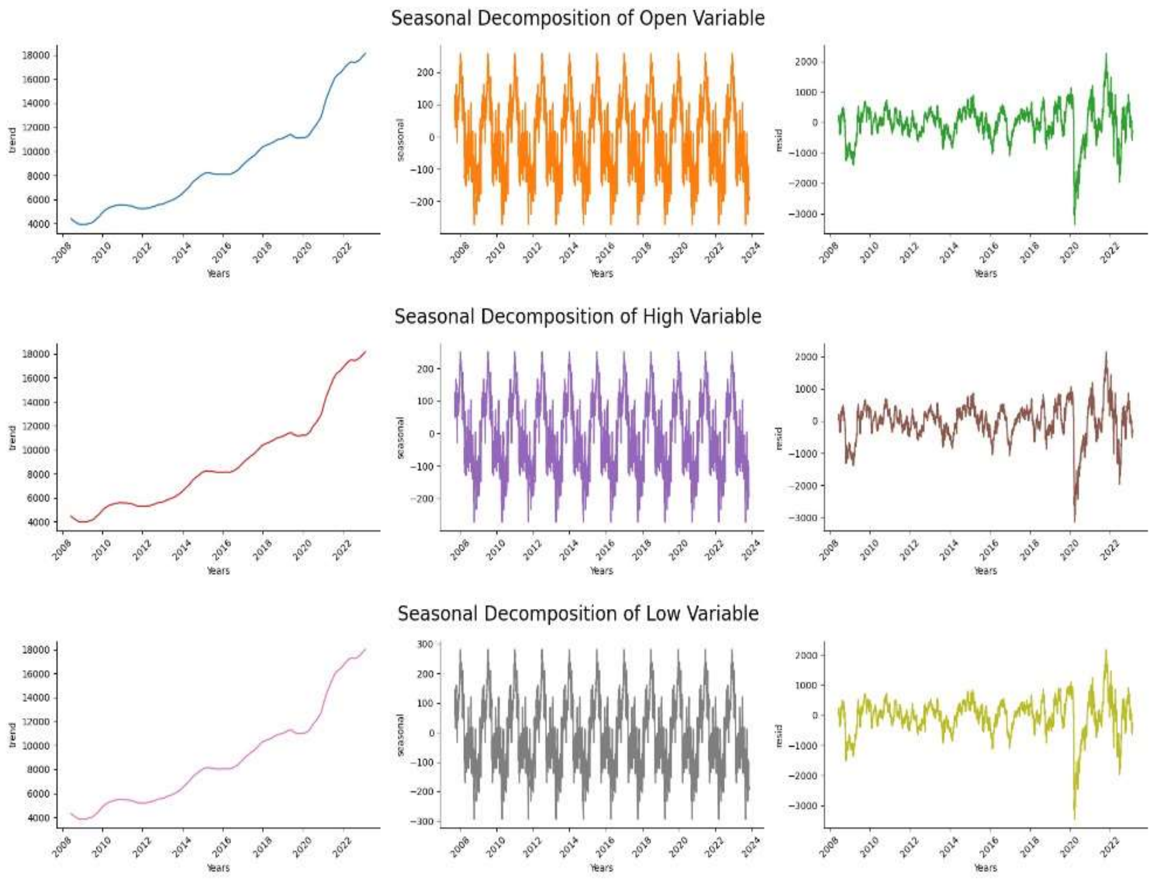
Variables in Dataset	Relative ordering test statistic value	Upper α-points of Normal distribution
Open	80.22	1.96
Close	80.23	1.96
Low	80.28	1.96
High	80.78	1.96

Hence, we can conclude that our null hypothesis is rejected at the level of significance 5% i.e. the trend component is obviously present in our data.

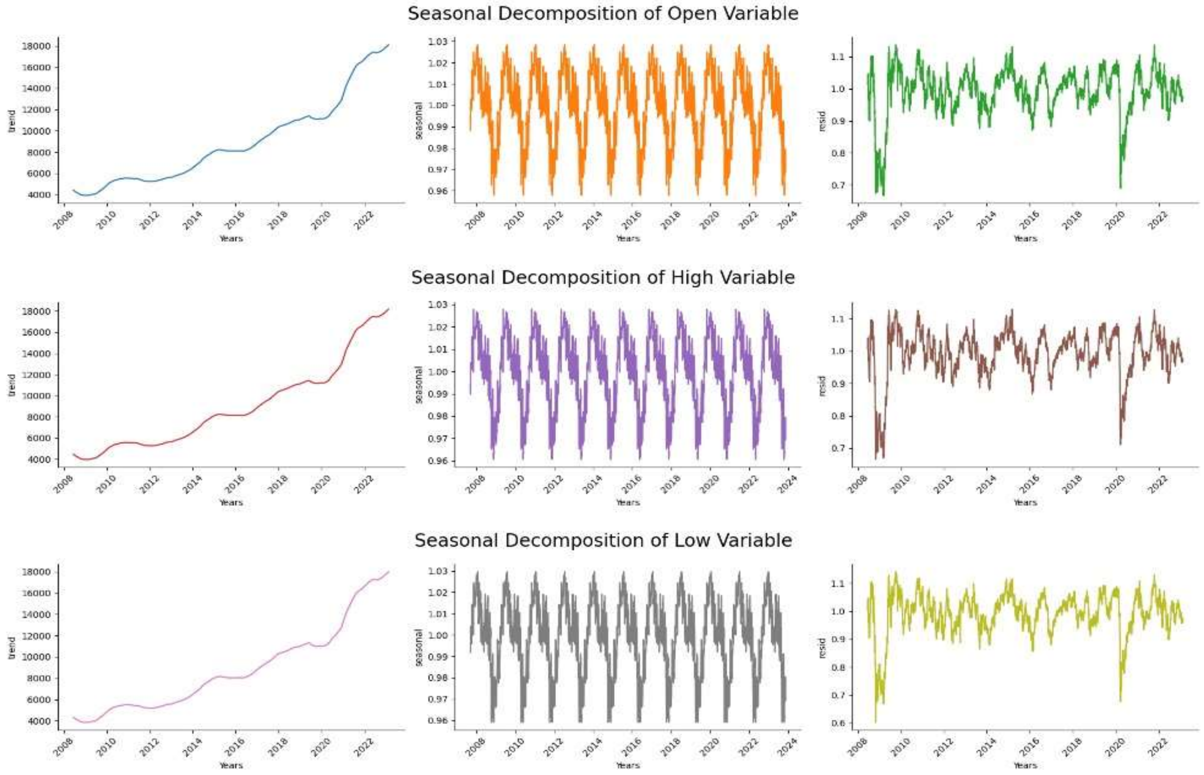
Seasonal Decompostion of the features:-

- Seasonal Decomposition of a series is a statistical process used to remove seasonal patterns from a time series dataset. This process involves breaking down the data into three components: the trend, the seasonal component, and the residuals.
- The trend component is a smooth representation of the series, the seasonal component is the characteristic pattern that repeats over time, and the residuals are the remaining variations in the data that are not explained by the trend and seasonal components.
- Seasonal Decomposition can be done using various techniques, such as moving average, smoothing, and exponential smoothing.

Under Additive model



Under multiplicative model



Conclusion : Not much of a qualitative change in trend and seasonality components, but the residuals looks much more stable around a constant level such phenomenon does not of course imply stationarity by itself, but at least a clear signal in the opposite direction is not there anymore.

Checking for Stationarity in the random part:

To select a suitable model for processed data, confirming stationarity is essential. Stationary data maintains consistent statistical properties, representing a stable state of statistical equilibrium. This ensures that the underlying system has settled into a steady state, preserving its statistical characteristics throughout the observed time period.

Test for Stationarity:

Stationarity check is a method of testing whether a time series data is stationary or not. It is done by plotting the data in a graph and checking for any clear trend and seasonality. Stationarity can be determined by looking at the mean, variance, and autocorrelation of the data.

- If a time series is stationary, the mean, variance, and autocorrelation should remain constant over time. If the mean, variance, or autocorrelation change over time, the time series is non-stationary.

Two statistical tests would be used to check the stationarity of a time series

Augmented Dickey Fuller (ADF) test

Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test.

Augmented Dickey-Fuller test (ADF):

An Augmented Dickey–Fuller test (ADF) is actually an augmented version of the Dickey–Fuller test used for large and more complicated version of time series models.
It is used to determine the presence of unit root in the series and hence helps in understand if the series is stationary or not. The ADF statistic is generally a negative number in the test. With the increase in negativity the rejection of hypothesis becomes fairly strong that is there is a unit root at some confidence level.

The null and alternate hypothesis of this test are:

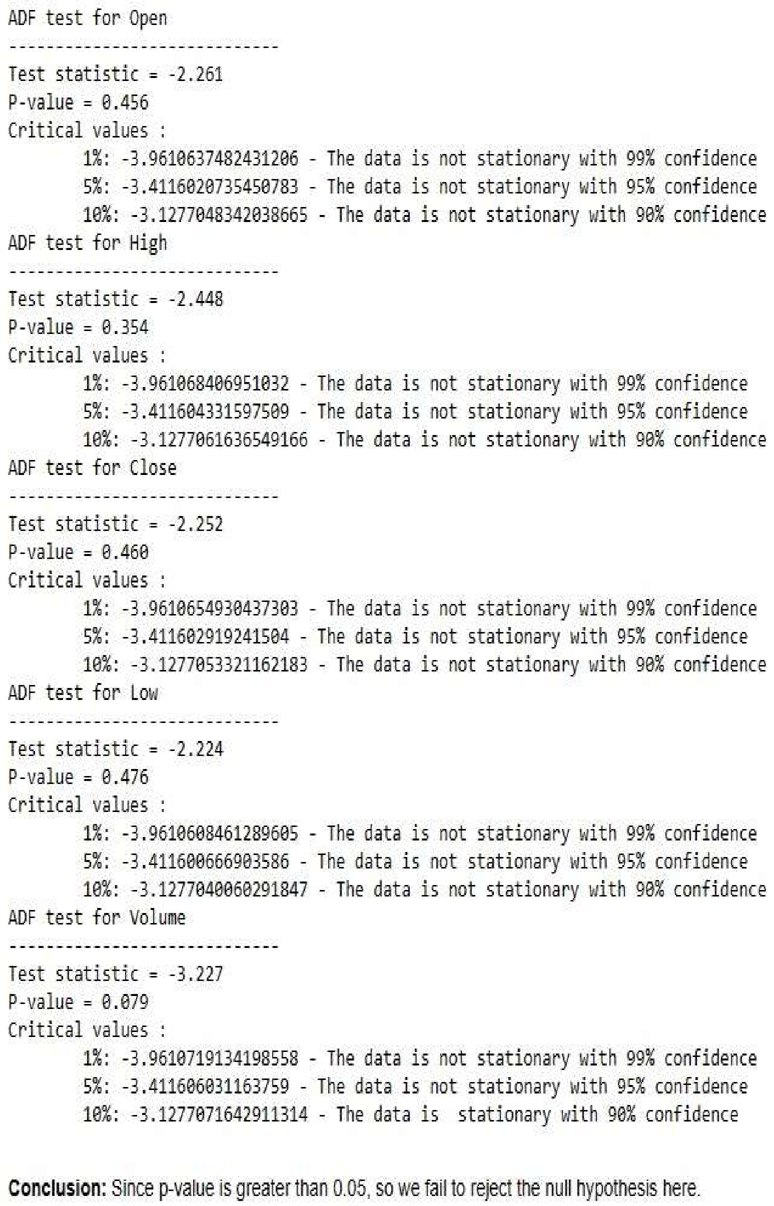
H_0: The series has a unit root.

against

H_1: The series has no unit root.

Conclusion: If the null hypothesis in failed to be rejected, this test may provide evidence that the series is non-stationary.

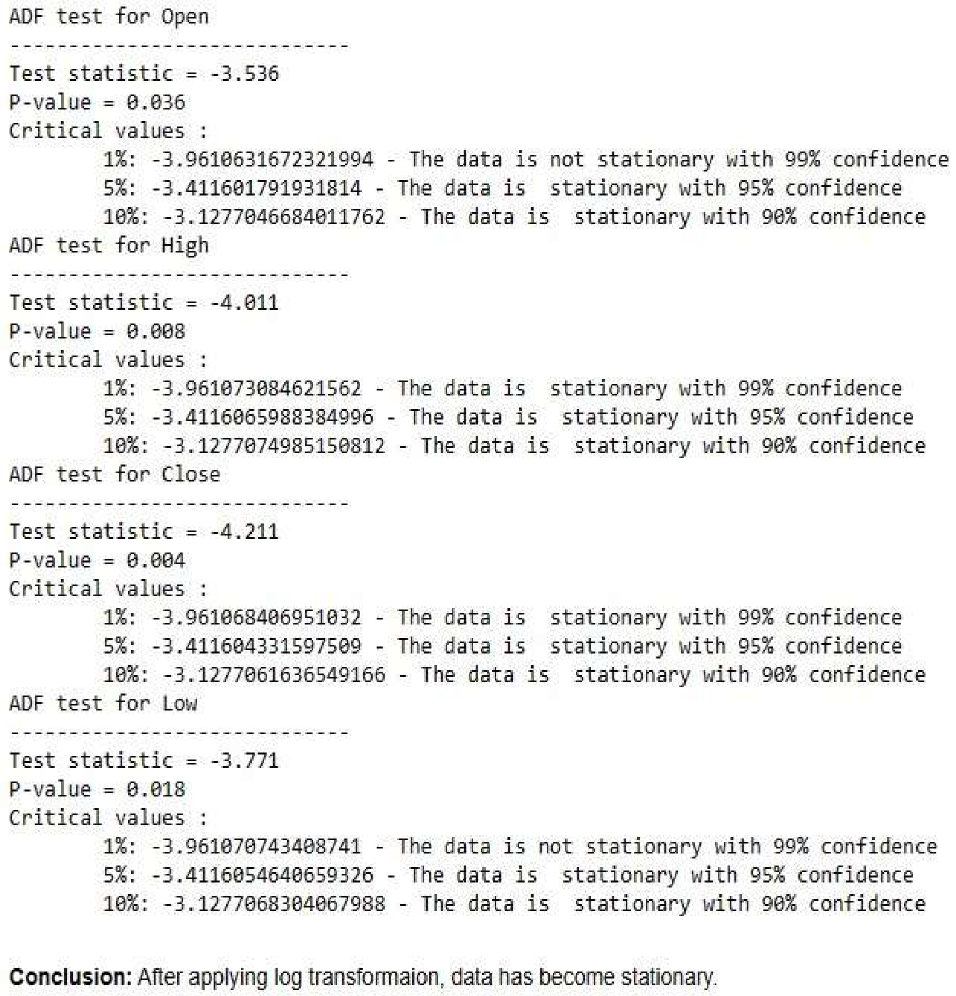
Output of ADF test



```
ADF test for Open
-----------------------------
Test statistic = -2.261
P-value = 0.456
Critical values :
        1%: -3.9610637482431206 - The data is not stationary with 99% confidence
        5%: -3.4116020735450783 - The data is not stationary with 95% confidence
        10%: -3.1277048342038665 - The data is not stationary with 90% confidence
ADF test for High
-----------------------------
Test statistic = -2.448
P-value = 0.354
Critical values :
        1%: -3.961068406951032 - The data is not stationary with 99% confidence
        5%: -3.411604331597509 - The data is not stationary with 95% confidence
        10%: -3.1277061636549166 - The data is not stationary with 90% confidence
ADF test for Close
-----------------------------
Test statistic = -2.252
P-value = 0.460
Critical values :
        1%: -3.9610654930437303 - The data is not stationary with 99% confidence
        5%: -3.411602919241504 - The data is not stationary with 95% confidence
        10%: -3.1277053321162183 - The data is not stationary with 90% confidence
ADF test for Low
-----------------------------
Test statistic = -2.224
P-value = 0.476
Critical values :
        1%: -3.9610608461289605 - The data is not stationary with 99% confidence
        5%: -3.411600666903586 - The data is not stationary with 95% confidence
        10%: -3.1277040060291847 - The data is not stationary with 90% confidence
ADF test for Volume
-----------------------------
Test statistic = -3.227
P-value = 0.079
Critical values :
        1%: -3.9610719134198558 - The data is not stationary with 99% confidence
        5%: -3.411606031163759 - The data is not stationary with 95% confidence
        10%: -3.1277071642911314 - The data is  stationary with 90% confidence
```

Conclusion: Since p-value is greater than 0.05, so we fail to reject the null hypothesis here.

Applying log transformation.



```
ADF test for Open
-----------------------------
Test statistic = -3.536
P-value = 0.036
Critical values :
        1%: -3.9610631672321994 - The data is not stationary with 99% confidence
        5%: -3.411601791931814 - The data is  stationary with 95% confidence
        10%: -3.1277046684011762 - The data is  stationary with 90% confidence
ADF test for High
-----------------------------
Test statistic = -4.011
P-value = 0.008
Critical values :
        1%: -3.961073084621562 - The data is  stationary with 99% confidence
        5%: -3.4116065988384996 - The data is  stationary with 95% confidence
        10%: -3.1277074985150812 - The data is  stationary with 90% confidence
ADF test for Close
-----------------------------
Test statistic = -4.211
P-value = 0.004
Critical values :
        1%: -3.961068406951032 - The data is  stationary with 99% confidence
        5%: -3.411604331597509 - The data is  stationary with 95% confidence
        10%: -3.1277061636549166 - The data is  stationary with 90% confidence
ADF test for Low
-----------------------------
Test statistic = -3.771
P-value = 0.018
Critical values :
        1%: -3.961070743408741 - The data is not stationary with 99% confidence
        5%: -3.4116054640659326 - The data is  stationary with 95% confidence
        10%: -3.1277068304067988 - The data is  stationary with 90% confidence
```

Conclusion: After applying log transformaion, data has become stationary.

After applying the log transformation we see that our data has become stationary.

KPSS test

It is a statistical test used to determine whether a time series is stationary around a deterministic trend. It is useful when there is suspicion that a time series might be trend-stationary rather than difference-stationary.

The null and alternate hypothesis of this test are:

H_0: The time series is stationary around a deterministic trend.
against
H_1: The time series has a unit root and is non-stationary.

In order to reject the null hypothesis, the test statistic should be greater than the provided critical values. If it is in fact higher than the target critical value, then that should automatically reflect in a low p-value.

That is, if the p-value is less than 0.05, the KPSS statistic will be greater than the 5% critical value.

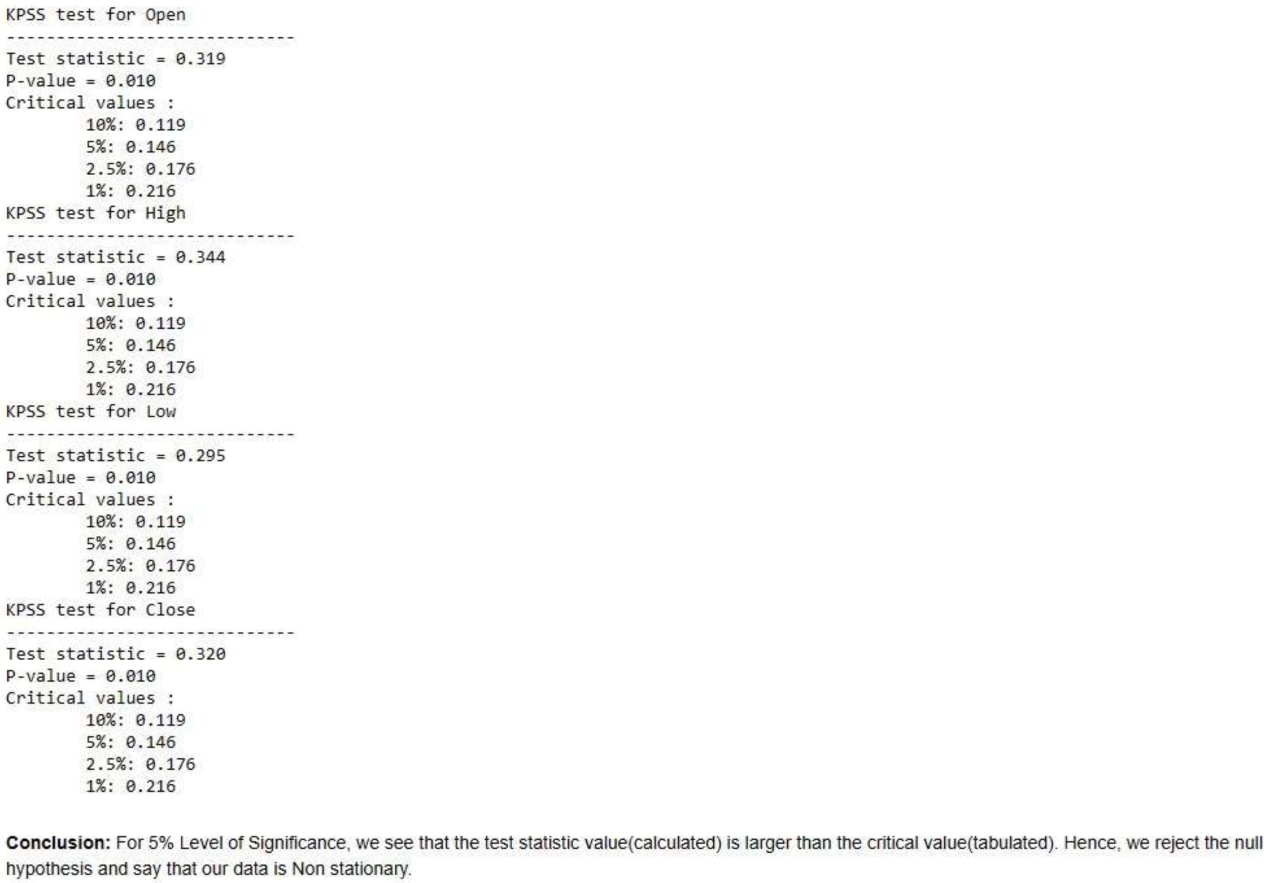
Finally, the number of lags reported is the number of lags of the series that was actually used by the model equation of the KPSS test.

By default, the statsmodels kpss() uses the ‘legacy’ method. In legacy method,

int(12 * (n / 100)**(1 / 4))

number of lags is included, where n is the length of the series.

Output



```
KPSS test for Open
-----------------------------
Test statistic = 0.319
P-value = 0.010
Critical values :
        10%: 0.119
        5%: 0.146
        2.5%: 0.176
        1%: 0.216
KPSS test for High
-----------------------------
Test statistic = 0.344
P-value = 0.010
Critical values :
        10%: 0.119
        5%: 0.146
        2.5%: 0.176
        1%: 0.216
KPSS test for Low
-----------------------------
Test statistic = 0.295
P-value = 0.010
Critical values :
        10%: 0.119
        5%: 0.146
        2.5%: 0.176
        1%: 0.216
KPSS test for Close
-----------------------------
Test statistic = 0.320
P-value = 0.010
Critical values :
        10%: 0.119
        5%: 0.146
        2.5%: 0.176
        1%: 0.216
```

Conclusion: For 5% Level of Significance, we see that the test statistic value(calculated) is larger than the critical value(tabulated). Hence, we reject the null hypothesis and say that our data is Non stationary.

ACF

The Autocorrelation Function (ACF) measures the correlation between a time series and its lagged values, illustrating how past observations influence present ones in a sequential manner.

PACF

Partial Autocorrelation Function (PACF) reveals the direct relationship between observations at different lags, providing insights into the unique contribution of each lag to the current value.

Assumption: Stationarity

- ACF and PACF assume stationarity of the underlying time series, so we have used log transformed data to calculate ACF and PACF.

Commonality in PCF and PACF

- Both the ACF and PACF start with a lag of 0, which is the correlation of the time series with itself and therefore results in a correlation of 1.
- The difference between ACF and PACF is the inclusion or exclusion of indirect correlations in the calculation.
- Additionally, you can see a blue area in the ACF and PACF plots.
- This blue area depicts the 95% confidence interval and is an indicator of the significance threshold.
- That means, anything within the blue area is statistically close to zero and anything outside the blue area is statistically non-zero.

For Integrated Order(d)

- Integrated order (I) is a parameter in an ARIMA model that refers to the number of times the data have been differenced in order to make it stationary.
- A value of 0 indicates that the data has not been differenced, while a value of 1 indicates that the data has been differenced once. Higher values indicate that the data has been differenced multiple times.

FOR MA(q)

- A moving average model is a type of time series model that uses historical data to forecast future values. It is a simple and widely used technique that can be used to smooth out short-term fluctuations in data and highlight longer-term trends and cycles.
- The model is based on the idea that the value of a time series at any given point is a function of the average of the values of the previous few points. The number of points used to calculate the average is called the window size.

FOR AR(p)

- An autoregressive (AR) model is a type of a statistical model used in time series analysis that uses observations from the past to predict future values.
- It is a type of regression analysis where the output variable (Y) is modelled as a linear function of its own past values and a noise term (epsilon).
- AR models are used to describe and analyse time-dependent phenomena such as stock prices, economic cycles, and ecological phenomena.

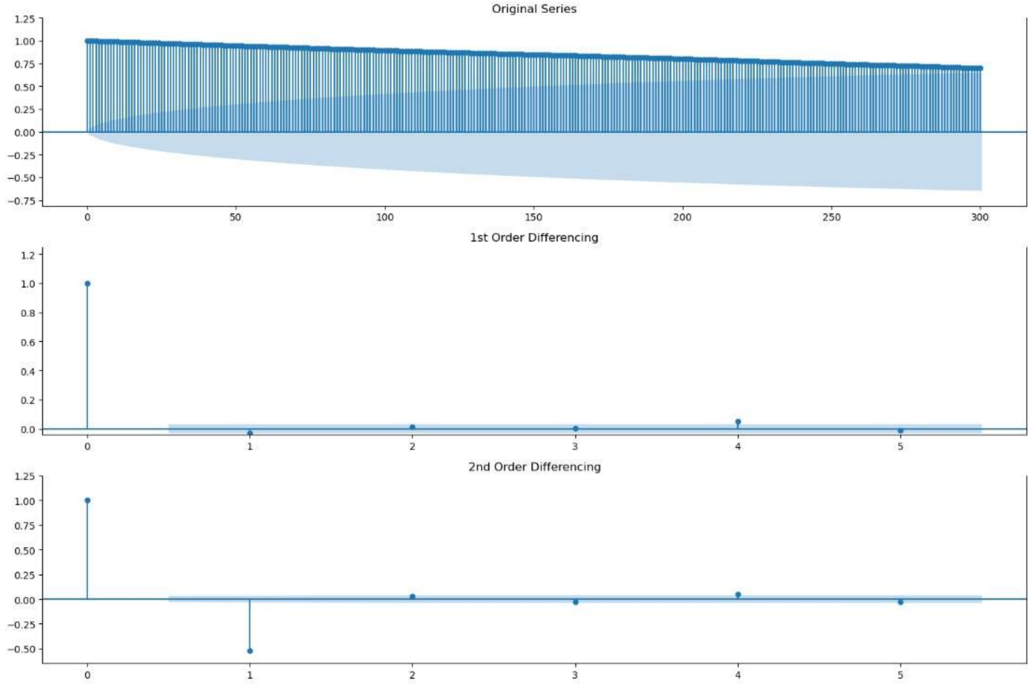
Finding out the orders for p and q

The p and q values in ARIMA are determined using a process called autocorrelation function (ACF) and partial autocorrelation function (PACF).

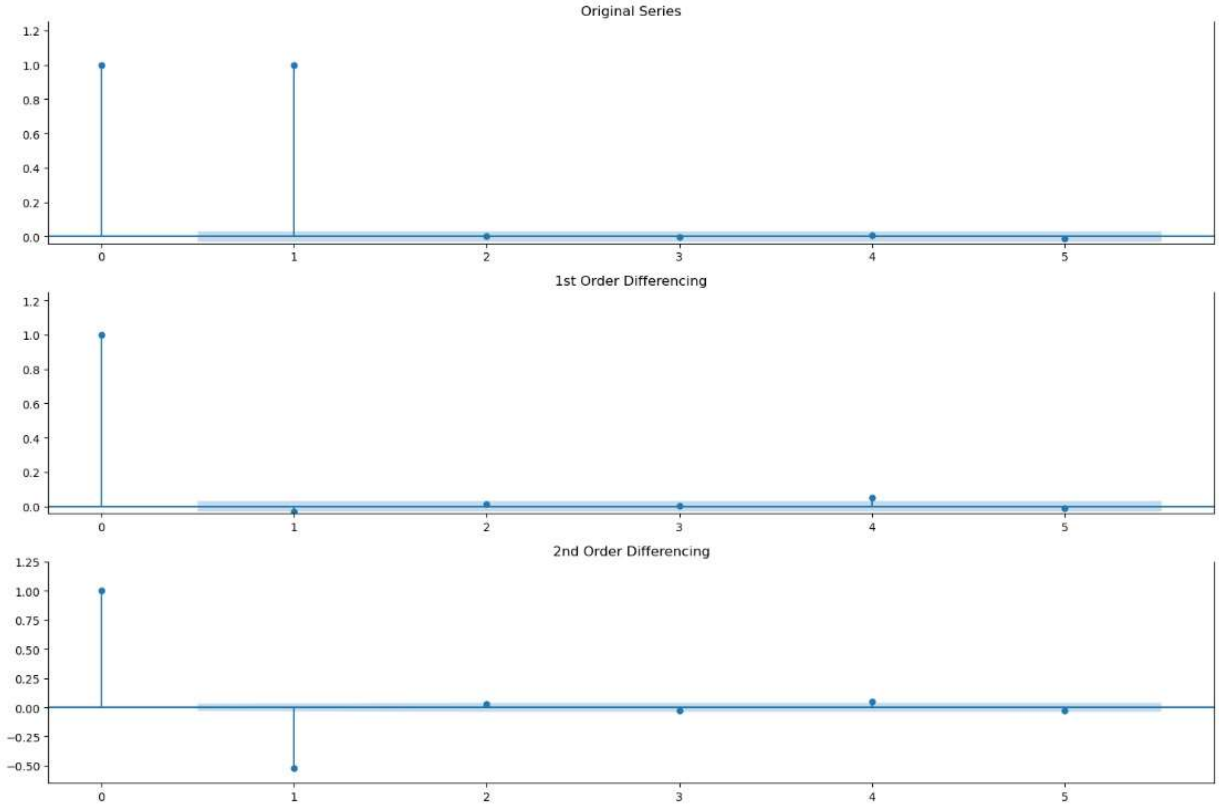
- The ACF and PACF help to identify the number of autoregressive (p) and moving average (q) terms needed in the model.
- Generally, if the ACF shows a sharp cutoff and the lag-1 autocorrelation is positive, we should use an AR model (p). If the PACF shows a sharp cutoff and the lag-1 autocorrelation is negative, we should use an MA model (q).
- If there is a gradual decline, we should use a combination of both (p and q). Finally, the order of the AR and MA model must be determined by minimizing the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC).

ACF/PACF PLOT for OPEN

ACF plot

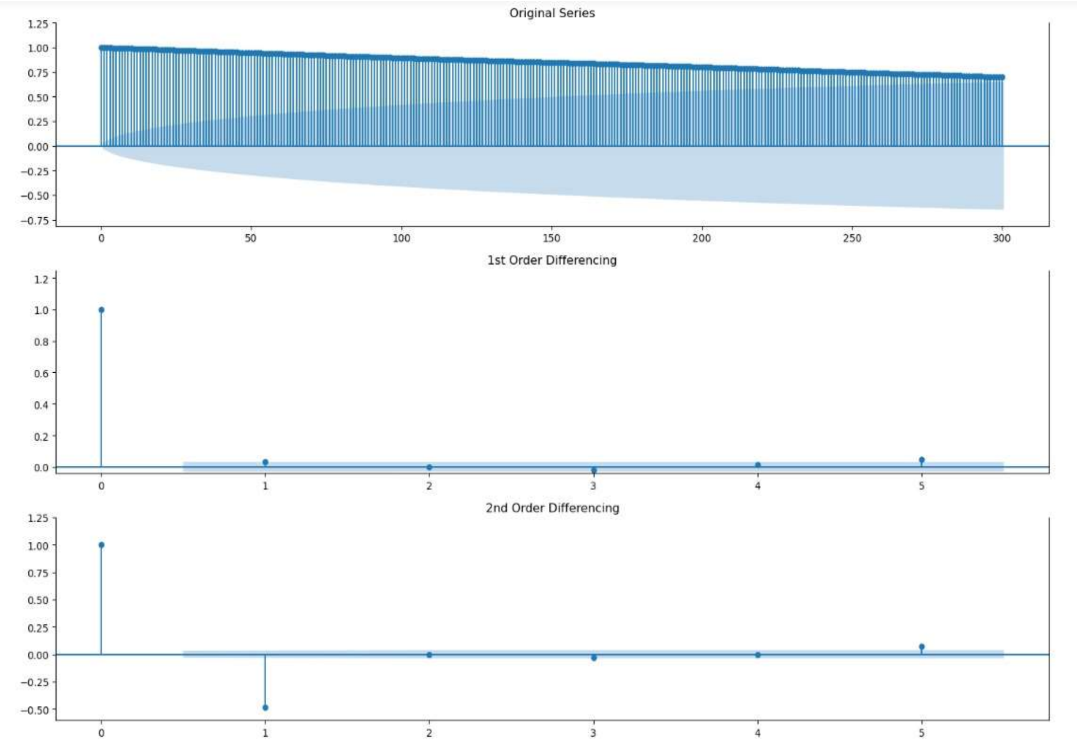


PACF plot

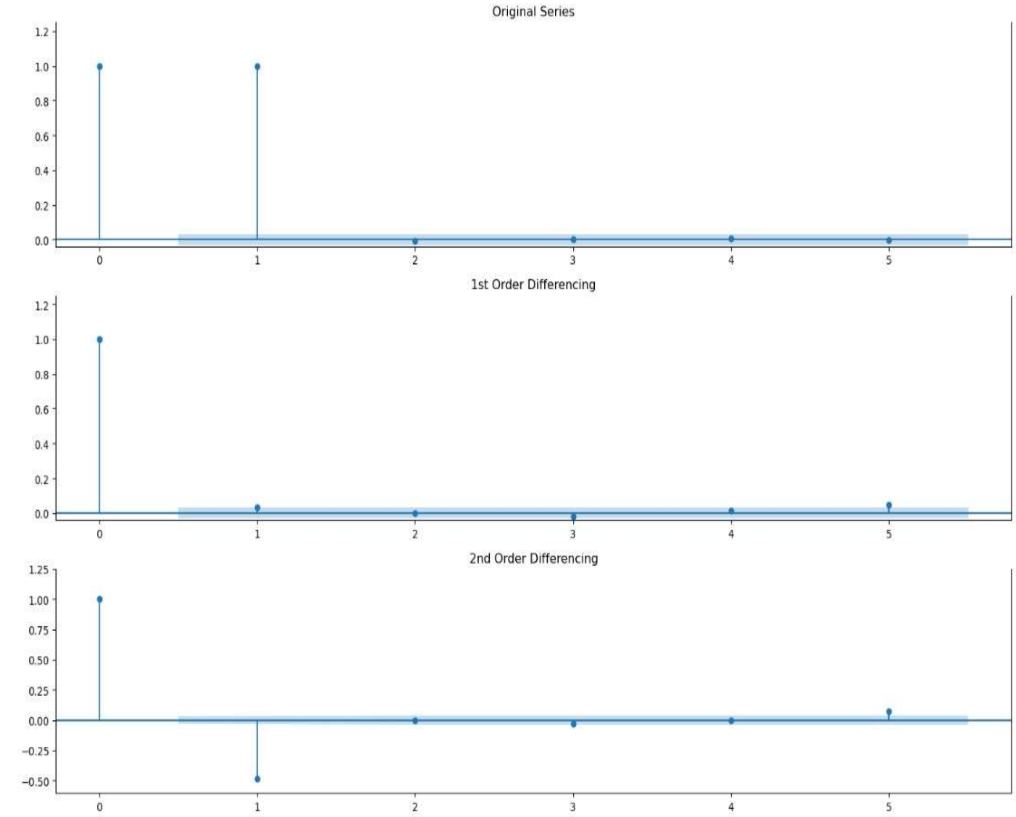


ACF/PACF PLOT for CLOSE

ACF plot

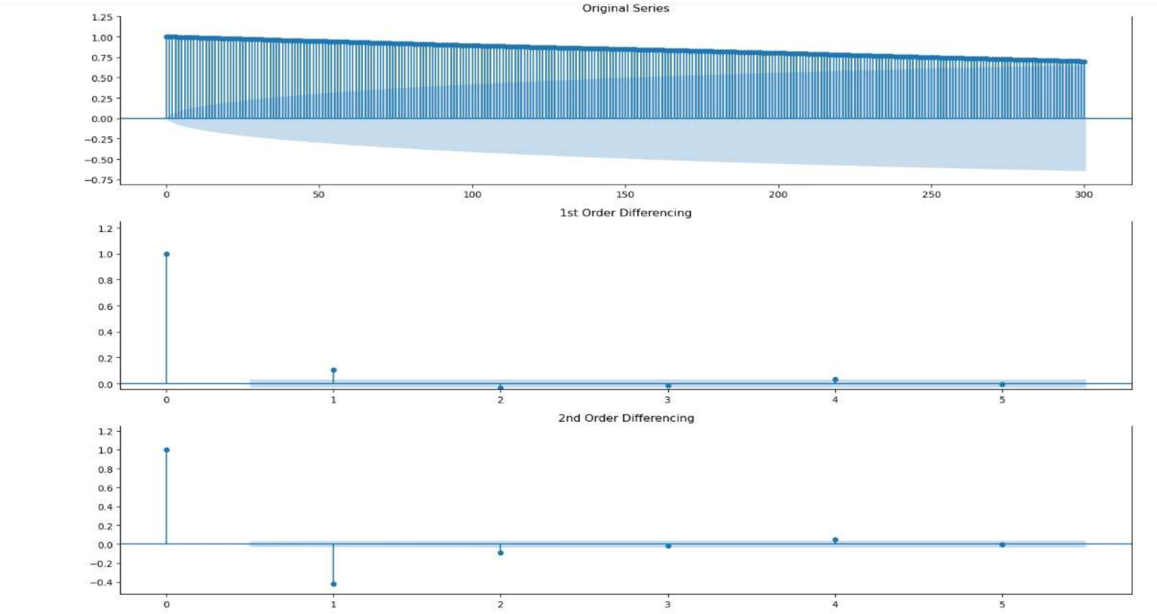


PACF plot

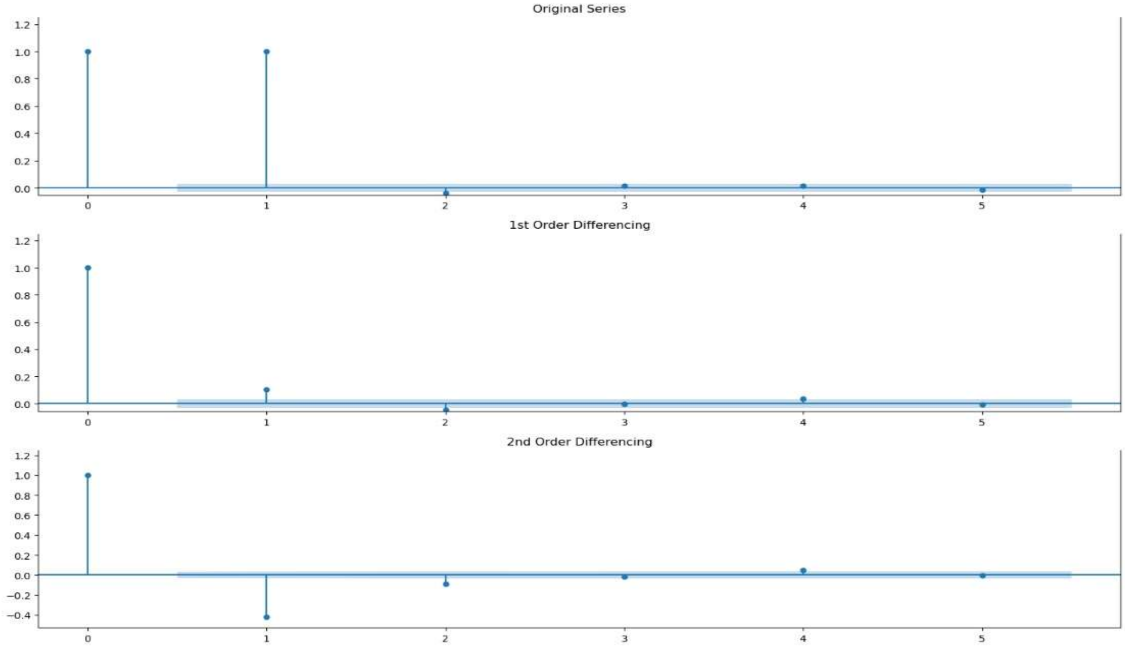


ACF/PACF PLOT for LOW

ACF plot :-

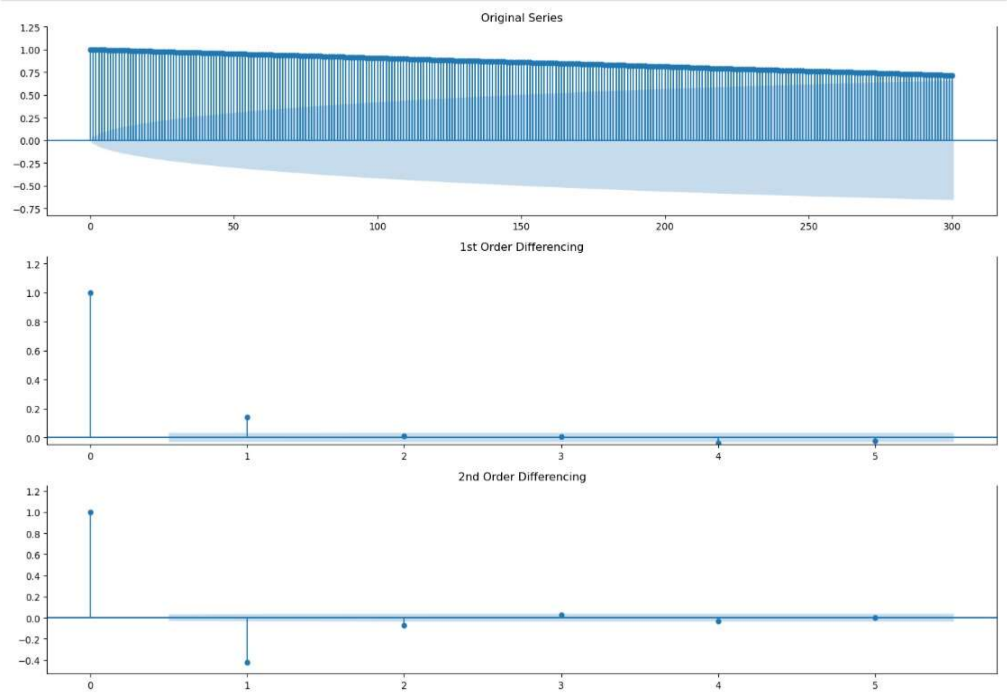


PACF plot :-

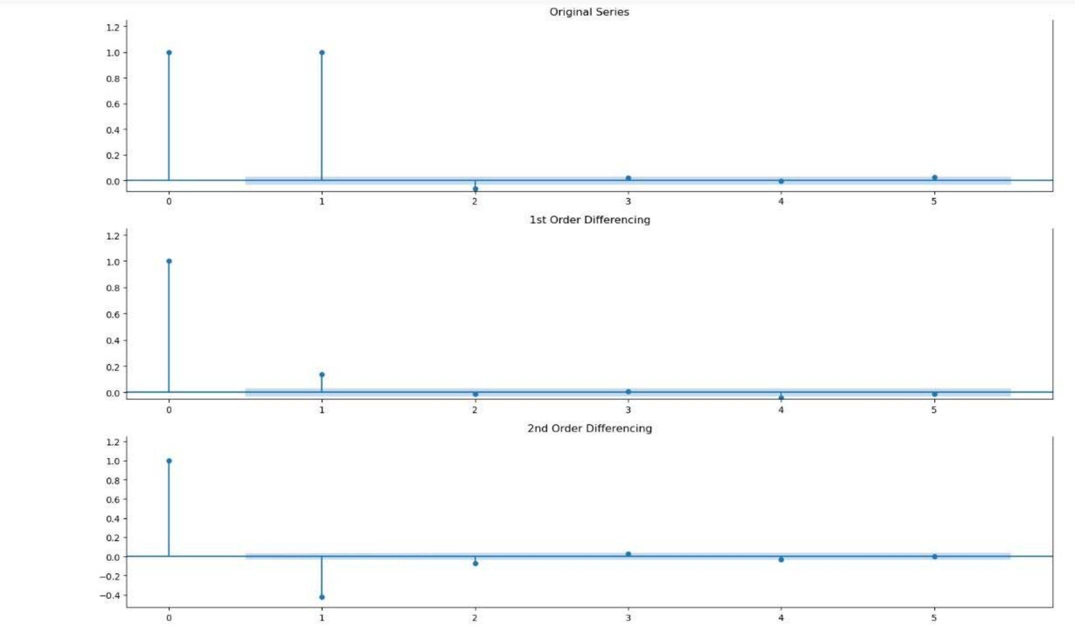


ACF/PACF PLOT for HIGH

ACF plot :-



PACF plot :-



ESTIMATING “p” AND “q”:

For estimating p and q we use the Akaike information criterion(AIC) and Bayesian information criterion (BIC).

Akaike information criterion(AIC)

For estimating p and q we use the Akaike information criterion.

The general form is given by AIC_k= -2logL(ϴ) + 2k where k is the number of paremeters in ϴ (vector of parameters).

For ARMA(p,q) $AIC_{p,q}$= 2logL(ϴ) + 2(p+q+1)

The pair (p̂, q̂) is obtained by minimizing $AIC_{p,q}$.

Table for p and q:

Features	p	q
Open	4	5
Close	3	3
Low	5	4
High	5	0

Bayesian Information Criterion (BIC)

The general form is given by BIC_k = -2 log L(ϴ) + k log(n), where k is the number of parameters in ϴ(vector of parameters), and n is the sample size.

For ARMA(p, q), BIC_p,q = log L(ϴ) + (p + q + 1) log(n).

The pair (p̂, q̂) is obtained by minimizing BIC_p,q.

Table for p and q:

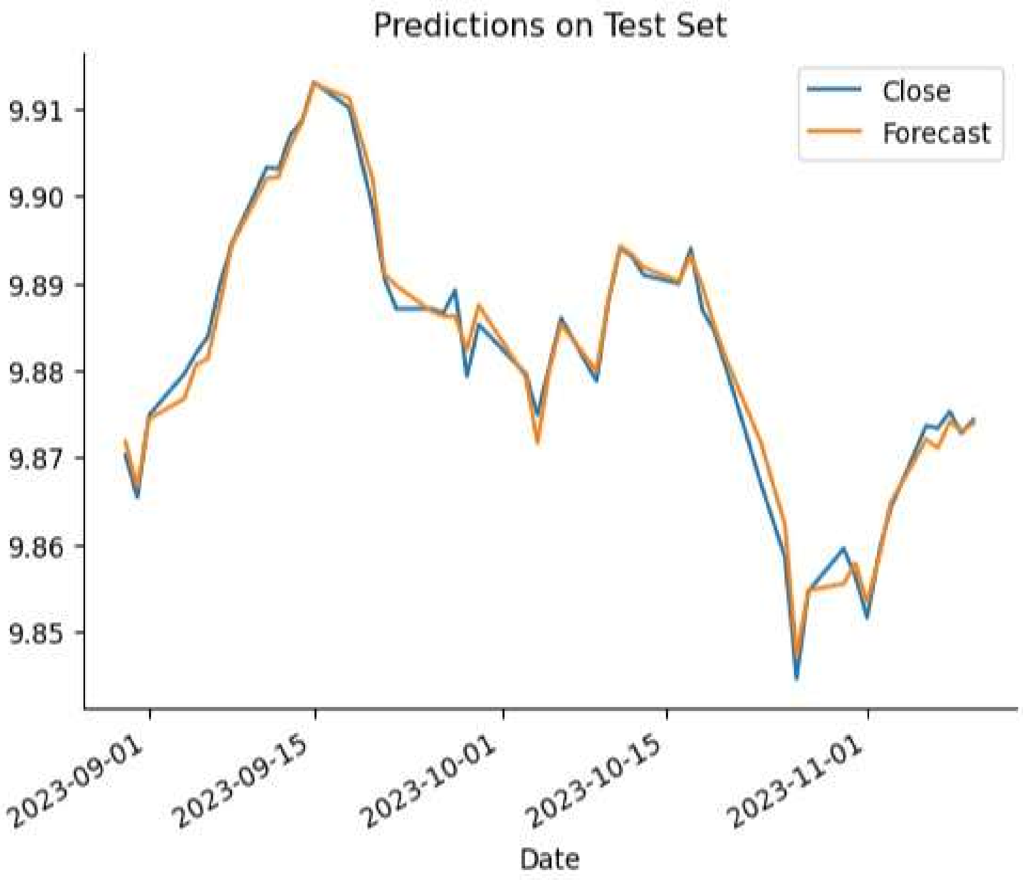
Features	p	q
Open	1	0
Close	1	1
Low	1	1
High	1	1

We will now go ahead and use Statistical Model ARIMA for making predictions. We will predict the Close Figure using the Open, High and Low Values.

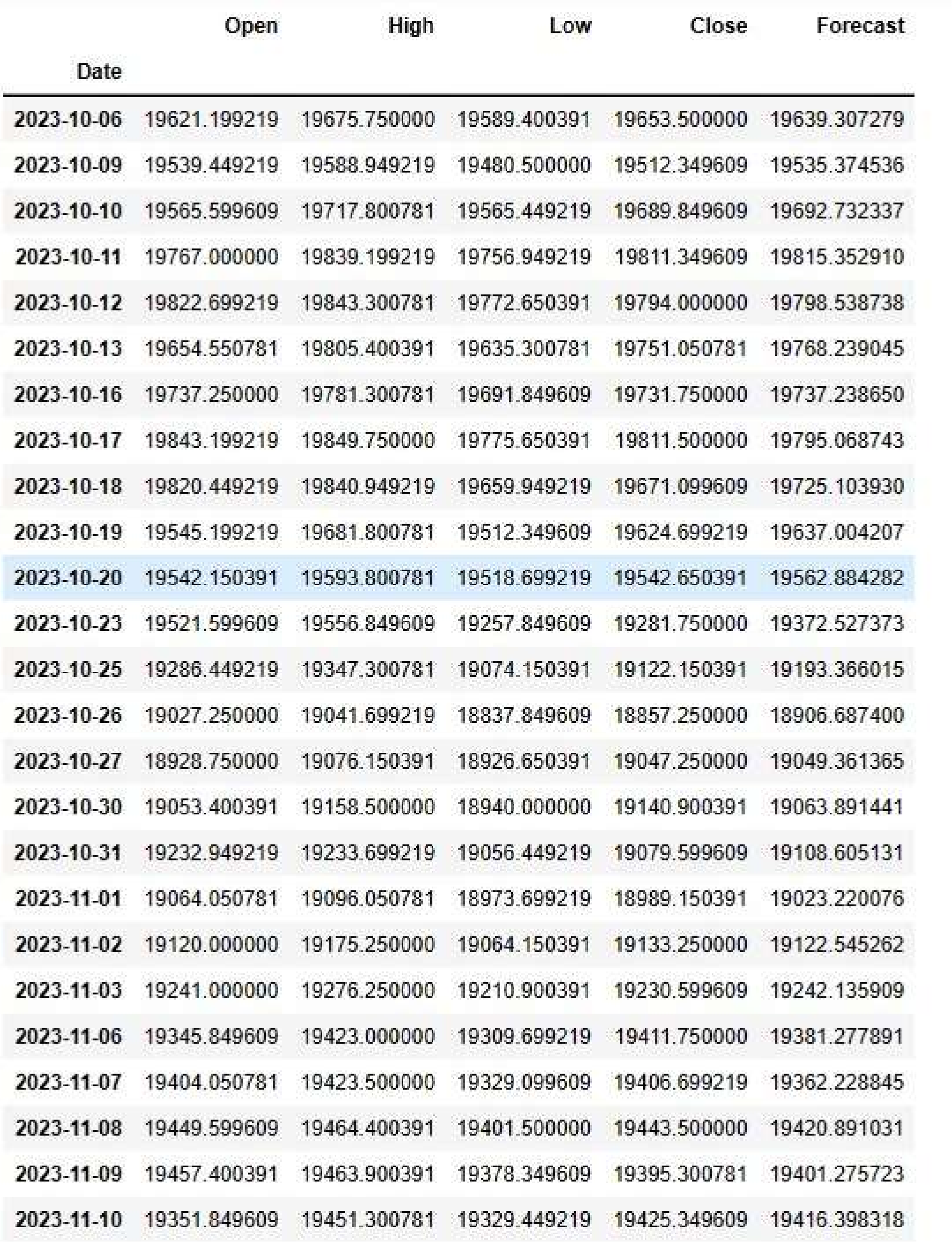
The value which we got for ARIMA(p,d,q) is (0,1,1)

Forecast

We forecasted for the close component.



The forecast of the close feature for the last 50 days is shown.



	Open	High	Low	Close	Forecast
Date					
2023-10-06	19621.199219	19675.750000	19589.400391	19653.500000	19639.307279
2023-10-09	19539.449219	19588.949219	19480.500000	19512.349609	19535.374536
2023-10-10	19565.599609	19717.800781	19565.449219	19689.849609	19692.732337
2023-10-11	19767.000000	19839.199219	19756.949219	19811.349609	19815.352910
2023-10-12	19822.699219	19843.300781	19772.650391	19794.000000	19798.538738
2023-10-13	19654.550781	19805.400391	19635.300781	19751.050781	19768.239045
2023-10-16	19737.250000	19781.300781	19691.849609	19731.750000	19737.238650
2023-10-17	19843.199219	19849.750000	19775.650391	19811.500000	19795.068743
2023-10-18	19820.449219	19840.949219	19659.949219	19671.099609	19725.103930
2023-10-19	19545.199219	19681.800781	19512.349609	19624.699219	19637.004207
2023-10-20	19542.150391	19593.800781	19518.699219	19542.650391	19562.884282
2023-10-23	19521.599609	19556.849609	19257.849609	19281.750000	19372.527373
2023-10-25	19286.449219	19347.300781	19074.150391	19122.150391	19193.366015
2023-10-26	19027.250000	19041.699219	18837.849609	18857.250000	18906.687400
2023-10-27	18928.750000	19076.150391	18926.650391	19047.250000	19049.361365
2023-10-30	19053.400391	19158.500000	18940.000000	19140.900391	19063.891441
2023-10-31	19232.949219	19233.699219	19056.449219	19079.599609	19108.605131
2023-11-01	19064.050781	19096.050781	18973.699219	18989.150391	19023.220076
2023-11-02	19120.000000	19175.250000	19064.150391	19133.250000	19122.545262
2023-11-03	19241.000000	19276.250000	19210.900391	19230.599609	19242.135909
2023-11-06	19345.849609	19423.000000	19309.699219	19411.750000	19381.277891
2023-11-07	19404.050781	19423.500000	19329.099609	19406.699219	19362.228845
2023-11-08	19449.599609	19464.400391	19401.500000	19443.500000	19420.891031
2023-11-09	19457.400391	19463.900391	19378.349609	19395.300781	19401.275723
2023-11-10	19351.849609	19451.300781	19329.449219	19425.349609	19416.398318