IIT KANPUR

Project Report - MTH517A

Time Series Analysis for Daily Stock Price Prediction of Indus Tower Limited

Arkaprova Saha (201278)
Bimal Roy (201292)
Rachita Mondal (201374)
Shreya Pramanik (201415)
Souvik Bhattacharyya (201433)

Supervisor: Prof. Amit Mitra

Acknowledgement

We want to express our thankfulness to our mentor Prof. Amit Mitra whose continuous support, guidance and lucid explanation of the topics have helped us to conduct work. We are truly grateful towards him for giving us a wonderful exposure to the topics of Time Series Analysis.

Abstract

The aim of our work is to apply relevant concepts of Time Series Analysis for daily closing stock price prediction of *Indus Tower Limited.* At first we tried to predict the mean behavior of the data set using appropriate AR, MA and ARIMA models. Thereafter, we have used Hybrid ARIMA-ARCH and Hybrid ARIMA-GARCH to capture the volatility of the series.

Contents

1 Introduction

Time Series is a series of observations recorded sequentially over a period of time. Data can be collected yearly, monthly, quarterly, weekly etc. Time Series Analysis has numerous application to various fields. It can be useful to analyse the data obtained from the economy, finance or medical field. It is used extensively for weather forecasting and digital signal processing problems. Identifying dominant components and explaining the time series through a random process are key features of a Time Series Analysis. We tried to conduct an appropriate time domain analysis on the observed data in this project.

2 Components Of a Time Series:

A time series is a collection of observations of well-defined data items obtained through repeated measurements over time. For example, measuring the value of retail sales each month of the year would comprise a time series. This is because sales revenue is well defined, and consistently measured at equally spaced intervals. Data collected irregularly or only once are not time series. An observed time series can be decomposed into three components: the trend (long term direction), the seasonal (systematic, calendar related movements) and the irregular (unsystematic, short term fluctuations).

2.1 Trend(Long Term Direction):

The trend shows the general tendency of the data to increase or decrease during a long period of time. A trend is a smooth, general, long-term, average tendency. It is not always necessary that the increase or decrease is in the same direction throughout the given period of time. If a time series does not show an increasing or decreasing pattern then the series is stationary in the mean.

2.2 Seasonal Component:

If there are regular and predictable fluctuations in the series that are correlated with the calendar - could be quarterly, weekly, or even days of the week, then the series includes a seasonality component. It's important to note that seasonality is domain specific, for example real estate sales are usually higher in the summer months versus the winter months while regular retail usually peaks during the end of the year. Also, not all time series have a seasonal component, as mentioned for audio or video data.

2.3 Cyclical Component:

Any pattern showing an up and down movement around a given trend is identified as a cyclical pattern. In a cyclical pattern the up and down movements do not occur in constant time intervals, they can not be predicted.

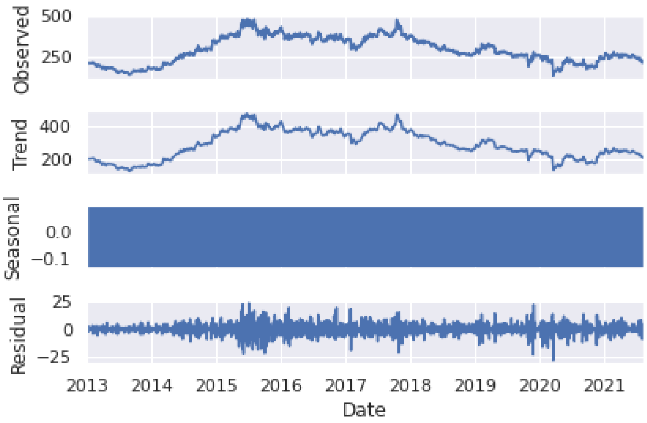
2.4 Random Component:

The residual is what's leftover when all the patterns have been removed. Residuals are random fluctuations. You can think of them as a noise component.

3 Decomposition Of a Time Series:

The decomposition of time series is a statistical task that deconstructs a time series into several components, each representing one of the underlying categories of pattern.Decomposition is the

deconstruction of the series data into its various components: trend, cycle, noise, and seasonality when those exist. Two different types of classic decomposition include multiplicative and additive decomposition.
The purpose of decomposition is to isolate the various components so we can view them each individually and perform analysis or forecasting without the influence of noise or seasonality.



4 Test for Stationarity:

A Stationary series is one whose statistical properties like mean, variance, covariance do not vary with time or these stats properties are not the function of time. In other words, stationarity in Time Series also means series without a Trend or Seasonal components.

4.1 Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test:

The KPSS test, short for, Kwiatkowski-Phillips-Schmidt-Shin (KPSS), is a type of Unit root test that tests for the stationarity of a given series around a deterministic trend.KPSS test is conducted with the following assumptions.

$$H_0 : \text{Series is trend stationary.}$$

$$H_1 : \text{Series is non-stationary.}$$

If the null hypothesis is failed to be rejected, this test may provide evidence that the series is trend stationary.
If Test statistic is less than Critical Value and p-value less than 0.05 – Reject Null Hypothesis(H_0) i.e., time series has a unit root, meaning it is not stationary. At 5% level p_value = 0.01 is less than 0.05, so we reject the Null Hypothesis , and conclude that the data is not stationary.

5 Test For Existence of Trend:

5.1 Mann-Kendall Test:

The non-parametric Mann-Kendall test is commonly employed to detect monotonic trends in series of environmental data, climate data or hydrological data. The null hypothesis, H_0, is that the data come from a population with independent realizations and are identically distributed. The alternative hypothesis, H_1, is that the data follow a monotonic trend. The Mann-Kendall test statistic is calculated according to :
The MK test is conducted as follows :

1. List the data in the order in which they were collected over time, $x_1, ...x_n$, which denote the measurements obtained at times 1,2,3...n respectively
2. Determine the sign of all n(n-1)/2 possible differences $x_j - x_k$, where $j \geq k$
These differences are: $x_2 - x_1, x_3 - x_1, \cdots, x_n - x_1, x_3 - x_2, x_4 - x_2, \cdots, x_n - x_{n-2}, x_n - x_{n-1}$

$$(x_j - x_k) = \begin{cases} 1 & \text{for } x_j - x_k \geq 0 \\ 0 & \text{for } x_j - x_k = 0 \\ -1 & \text{for } x_j - x_k \leq 0 \end{cases} \tag{1}$$

For example, if $x_j - x_k \geq 0$, that means that the observation at time j,denoted by x_j, is greater than the observation at time k, denoted by x_k. S= $\sum\limits_{j=k+1}^{n} \sum\limits_{k=1}^{n-1} (x_j - x_k)$ E(S) = 0 and the variance σ^2 is found to be

$$\sigma^2 = \left(n(n-1)(2n+5) - \sum_{j=1}^{p} t_j (t_j - 1)(2t_j + 5) \right) /18$$

If the observed value of S is significantly different from the expected value, that indicates the presence of trend. The form of Z is

$$Z = \begin{cases} \frac{S-1}{\sigma}, & \text{for } S \geq 0 \\ 0, & \text{for } S = 0 \\ \frac{S+1}{\sigma}, & \text{for } S \leq 0 \end{cases}$$

And S is related to Kendall's τ as

$$\tau = \frac{S}{D}$$

and

$$D = \left(n(n-1)/2 - \frac{1}{2} \sum_{j=1}^{p} t_j (t_j - 1) \right)^{\frac{1}{2}} (n(n-1)/2)^{\frac{1}{2}}$$

Asymptotic Test for H_0: No trend at level of observed $|Z| > \tau_{\frac{\alpha}{2}}$(upper),

$$Z = \frac{\tau - E[\tau]}{\sqrt{\tau}} \overset{asym}{\sim} N(0,1)$$

under H_0 as $n \to \infty$. We could reject the null hypothesis of no trend as level of significance α if observed $|Z| > \tau_{\alpha_2}$

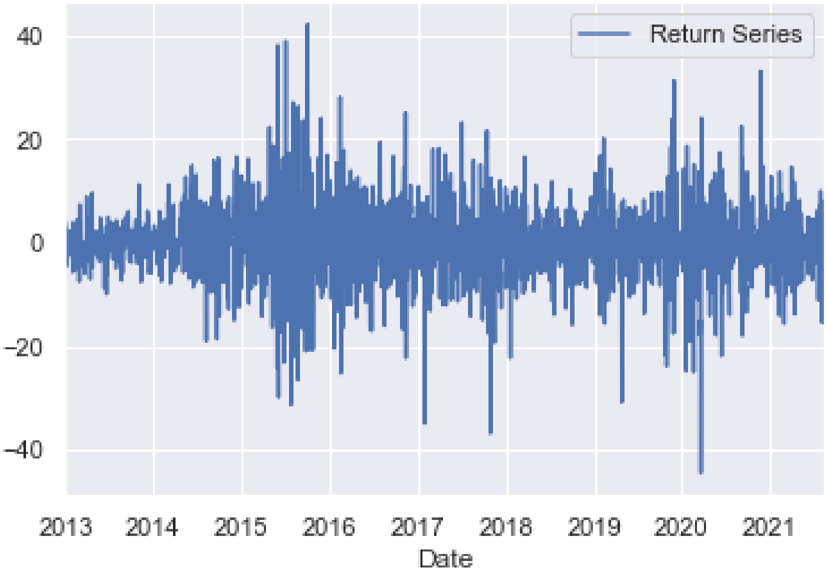
Conclusion: Depending upon the statistic we can determine that p-value is small, so we reject the Null hypothesis and conclude that trend is significant in the data.

6 Elimination of Deterministic component:

From the plot of the original time series data we suspect presence of trend component in the dataset. Hence, first order differencing on the data is performed for elimination of trend and We can see that

$$Z_t = \frac{(Y_t - Y_{t-1})}{Y_t}$$

is more or less stationary. After performing first order difference operation, we perform the kpss test on the return series data and find the p_value to be 0.1 which is greater than 0.05. Thus we can say that we fail to reject the null hypothesis and conclude that time series is stationary.



7 Introduction: MA Model

In time series analysis, the moving-average model (MA model), also known as moving-average process, is a common approach for modeling univariate time series. The moving-average model specifies that the output variable depends linearly on the current and various past values of a stochastic (imperfectly predictable) term. Because of the sequential nature of our time-series data, we need a way to aggregate this sequence of information. From all the potential techniques, the most simple one is the Moving Average (MA) Model.

7.1 Definition

The moving average model of order q (MA(q)) is defined as

$$X_t = \mu + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + ... + \theta_q\epsilon_{t-q}$$

where μ is the mean of the series, the $\theta_1, \theta_2, .., \theta_q$ are the parameters of the model and the $\epsilon_{t-1}, \epsilon_{t-2}, .., \epsilon_{t-q}$ are white noise error terms. The value of q is called the order of the MA model. This can be equivalently written in terms of the back-shift operator B as

$$X_t = \mu + (1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q)\epsilon_t.$$

Thus, a moving-average model is conceptually a linear regression of the current value of the series against current and previous (observed) white noise error terms or random shocks. The random shocks at each point are assumed to be mutually independent and to come from the same distribution, typically a normal distribution, with location at zero and constant scale.

7.2 Fitting the model

Fitting the MA estimates is more complicated because the lagged error terms are not observable. This means that iterative non-linear fitting procedures need to be used in place of linear least squares. For an MA(q) process, the autoorrelation function (ACF) is given by

$$\rho_X(h) = \begin{cases} 1 & \text{if h} = 0 \\ \frac{\sum_{j=0}^{q-|h|}\theta_j\theta_{j+h}}{(1+\sum_{j=1}^{q}\theta_j^2)} & \text{if } |h| \leq q \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The autocorrelation function (ACF) of an MA(q) process is zero at lag q + 1 and greater. Therefore, we determine the appropriate maximum lag for the estimation by examining the sample autocorrelation function to see where it becomes insignificantly different from zero for all lags beyond a certain lag, which is designated as the maximum lag q.

7.2.1 Determination of Confidence Interval for ACF

Let us suppose that the sample ACF for MA process is given by $\hat{\rho}(h)$, for lag h. Our purpose is to test whether the appropriate order is q. Hence the hypothesis of interest is :

$$\begin{aligned} H_0 &: \text{Model is of order q} \\ H_1 &: \text{Model Order is greater than q} \end{aligned}$$

Now under under null hypothesis we would expect that $\rho_X(h) = 0$, for $h > q$. Hence testing the above hypothesis is equivalent to testing:

$$\begin{aligned} &H_0 : \rho_X(h) = 0 \\ &H_1 : \rho_X(h) > 0 \qquad\qquad\qquad \text{for all } \ h > q \end{aligned}$$

For large n, $\hat{\rho}(h)$ follows $Normal(0, \frac{1}{n})$, where n is the sample size. In our case $n = 2128$. So unde 95% level of significance we would reject the null hypothesis if,

$$|\hat{\rho}(h)\ | > \frac{1.96}{\sqrt{n}} = 0.04248$$

Therefore, if any $\hat{\rho}(h) > 0.04248$, we can take that $\rho_X(h)$ to be significantly greater than 0 and hence the confidence band is given by $[-0.04248, 0.04248]$. And if our true model is of order q then we would expect that the values of $\hat{\rho}(h)$ would lie inside the confidence band for all $h > q$.

7.2.2 Akaike Information Criterion (AIC)

The general form of AIC is:

$$\text{AIC(k)} = -2\log\hat{L} + 2k,$$

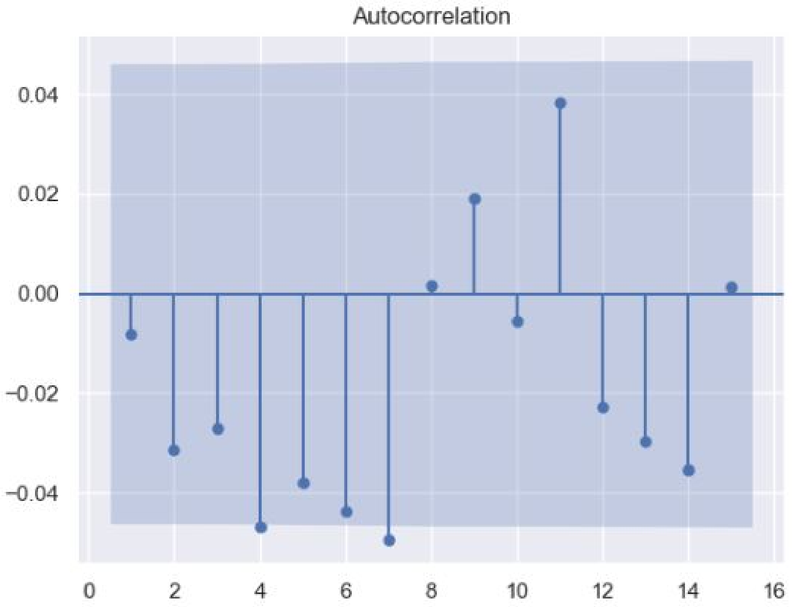
where k is number of parameters in the model. For MA(q) model

$$\text{AIC(q)} = -2\log\hat{L} + 2(q+1),$$

and $\hat{p} = \underset{p\in\{0,1,\ldots,p\}}{\arg\min}\ \text{AIC(q)}$.

7.3 Model Building

The ACF plot of return by using the statsmodels module of python with lag = 15 is as follows:



From the plot, we see a "spike" at lag 4 and at lag 7 followed by generally non-significant values for other points. So we will start by fitting MA model of order 4 and continue up-to fitting that of order 8 and select the model for which AIC is minimum.

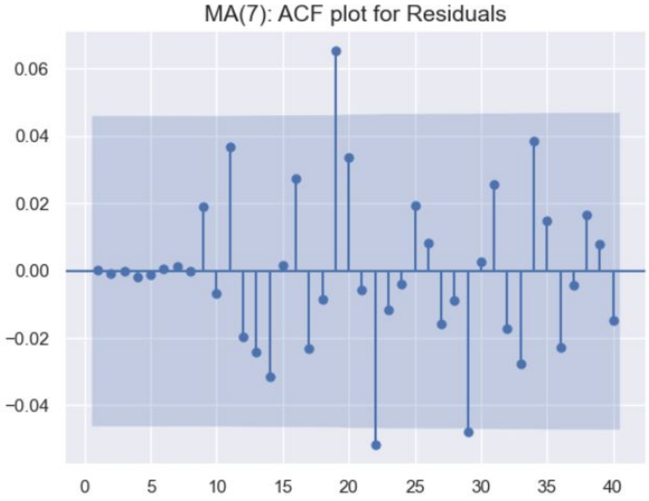
AIC scores with the corresponding models are given below:

Model	AIC
MA(4)	7838.922
MA(5)	7837.693
MA(6)	7836.100
MA(7)	7833.385
MA(8)	7835.378

And the p-values of the estimates of different lags for an MA(8) model are as follows:

Lag	Estimate	p-value
1	-0.0182	0.439
2	-0.0377	0.109
3	-0.0327	0.169
4	-0.0477	0.046
5	-0.0399	0.098
6	-0.0473	0.052
7	-0.0536	0.030
8	-0.0020	0.933

As we can see at the MA(8) output, the AIC value is starting to increase and the corresponding p-value is also high for lag 8. So we stick to MA(7) model.
Now let us check the ACF plot for the residuals of the MA(7) model.



Though majority of points lies inside the blue area, the existence of some points outside the area indicates that there exists a better predictor.

8 Introduction: AR Model

In statistics, econometrics and signal processing, an autoregressive (AR) model is a representation of a type of random process; as such, it is used to describe certain time-varying processes in nature, economics, etc. The autoregressive model specifies that the output variable depends linearly on its own previous values and on a stochastic term (an imperfectly predictable term); thus the model is in the form of a stochastic difference equation (or recurrence relation which should not be confused with differential equation).

8.1 Definition

An autoregressive model of order p or AR(p) is defined as

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t,$$

where $\phi_1, \phi_2, ..., \phi_p$ are the parameters of the model, c is a constant, and ϵ_t is white noise. This can be equivalently written using the backshift operator B as,

$$X_t = c + \sum_{i=1}^{p} \phi_i B^i X_t + \epsilon_t,$$

so that, moving the summation term to the left side and using polynomial notation, we have

$$\phi[B] X_t = c + \epsilon_t.$$

8.2 Fitting the model

8.2.1 Determination of Confidence Band for PACF

Let us suppose that the sample PACF for AR process is given by $\hat{\alpha}(h)$, for lag h. Our purpose is to test whether the appropriate order is p. Hence the hypothesis of interest is :

$$H_0 : \text{Model is of order p}$$
$$H_1 : \text{Model Order is greater than p}$$

Now under under null hypothesis we would expect that $\alpha_X(h) = 0$, for $h > p$. Hence testing the above hypothesis is equivalent to testing:

$$\begin{aligned} &H_0 : \alpha_X(h) = 0 \\ &H_1 : \alpha_X(h) > 0 \qquad\qquad\qquad \text{for all } \ h > p \end{aligned}$$

For large n, $\hat{\alpha}(h)$ follows $Normal(0, \frac{1}{n})$. So under 95% level of significance we would reject the null hypothesis if,

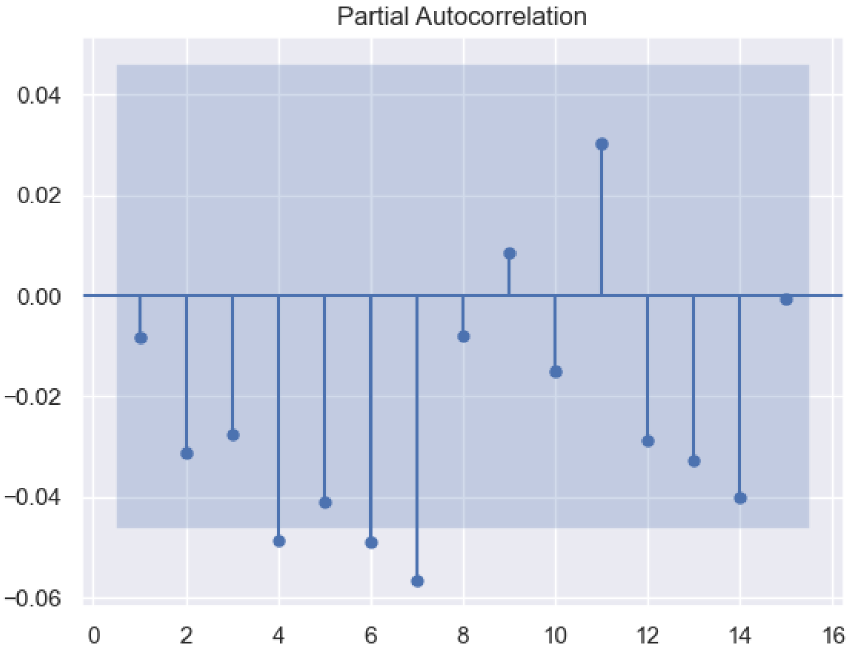
$$|\hat{\alpha}(h)\ | > 0.04248$$

Therefore, if any $\hat{\alpha}(h) > 0.04248$, we can take that $\alpha_X(h)$ to be significantly greater than 0 and hence the confidence band is given by $[-0.04248, 0.04248]$. And if our true model is of order p then we would expect that the values of $\hat{\alpha}(h)$ would lie inside the confidence band for all $h > p$.

8.2.2 Order Selection

The autocorrelation function of MA(q) models is zero for all lags greater than q as these are q-correlated processes. Hence, the ACF is a good indication of the order of the process. However AR(p) and ARMA(p,q) processes are "fully" correlated, their ACF tails off and never becomes zero, though it may be very close to zero. In such cases it is difficult to identify the process on the ACF basis only.

Looking at PACF (Partial Autocorrelation function) plot can help us form an idea of what the parameter of AR model might be. The PACF plot of return by using the statsmodels module of python with lag = 15 is as follows:



As we can see from the plot lag 4, lag 5 and lag 7 shows significant pacfs. So we will consider them and choose the model with lesser AIC and higher Log likelihood.

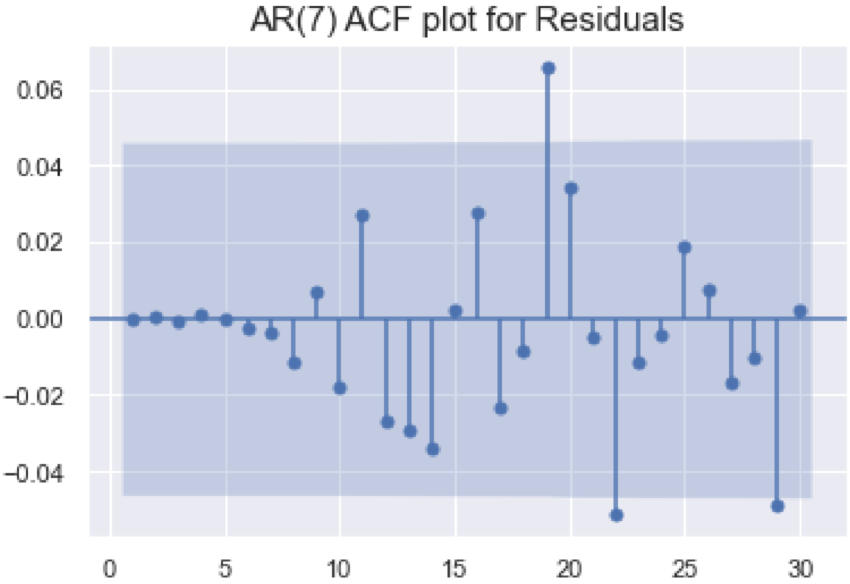
AIC scores with the corresponding models are given below:

Model	AIC
AR(4)	7841.066
AR(5)	7840.057
AR(6)	7837.629
AR(7)	7833.888
AR(8)	7835.785

And the p-values of the estimates of different lags for an AR(8) model are as follows:

Lag	Estimate	p-value
1	-0.0178	0.451
2	-0.0392	0.096
3	-0.0338	0.151
4	-0.0532	0.025
5	-0.0447	0.060
6	-0.0509	0.033
7	-0.0573	0.016
8	-0.0077	0.749

As we can see at the AR(8) output, the AIC value is starting to increase and the corresponding p-value is also high for lag 8. So we stick to AR(7) model.
Now let us check the ACF plot for the residuals of the AR(7) model.



Though majority of points lies inside the blue area. The existence of some points outside the area indicates that there might exist a better predictor.

9 Introduction: ARIMA

In statistics and econometrics, in particular in time series analysis, an autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting).ARIMA models are applied in some cases where data show evidence of non-stationarity, where an initial differencing step can be applied one or more times to eliminate the non-stationarity.

9.1 Explanation

The AR part of ARIMA indicates that the evolving variable of interest is regressed on its own lagged values. The MA part indicates that the regression error is actually a linear combination of error terms whose values occurred contemporaneously and at various times in the past. The I (for "integrated") indicates that the data values have been replaced with the difference between their values and the previous values (and this differencing process may have been performed more than once).

9.2 Definition

$\{X_t\}$ is said to follow an Auto Regressive Integrated Moving Average (ARIMA) model of order (p,d,q) if,

$$Z_t = \nabla^d X_t = (1-B)^d X_t \sim \text{ARMA}(p,q)$$

where ∇ = (1-B) and B is the difference operator and B is the back-shift operator, so it follows,

$$Z_t = \phi_1 Z_{t-1} + \cdots + \phi_p Z_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$$

where $Z_t = \nabla^d X_t$ i.e. d^{th} difference of X_t and $\phi_1, \phi_2, \cdots, \phi_p$ are the MA parameters of the model, and $\theta_1, \theta_2, \cdots, \theta_p$ are the AR parameters, and $\epsilon_t, \epsilon_{t-1}, \cdots, \epsilon_{t-q}$ are white noise error terms. This can as well be written in terms of back-shift operator as,

$$\phi(B)Z_t = \theta(B)\epsilon_t \implies \phi(B)\nabla^d X_t = \theta(B)\epsilon_t \implies \phi(B)(1-B)^d X_t = \theta(B)\epsilon_t$$

$$\phi^*(B)X_t = \theta(B)\epsilon_t \tag{3}$$

$$\text{Where } \phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p \text{ and } \theta(B) = 1 - \theta_1 B + \cdots + \theta_q B^q$$

$$\text{Thus from(3) the model for } \{X_t\} \text{ is } \phi^*(B)X_t = \theta(B)\epsilon_t$$

$$\phi^*(B) = \phi(B)(1-B)^d \implies X_t \sim \text{ARMA}(p+d,q)$$

$$\text{So, ARIMA(p,d,q)} \implies \text{ARMA(p+d,q)}$$

9.3 Model Fitting

Using the pdarima model we found out that ARIMA(2,0,1) is the best fit for our data. AIC for some of the corresponding ARIMA models are given below:

ARIMA Model	AIC
(2,0,2)	7830.881
(0,0,0)	7840.621
(1,0,0)	7842.498
(0,0,1)	7842.490
(0,0,0)	7839.047
(1,0,2)	7844.295
(2,0,1)	7830.781
(1,0,1)	7831.239
(2,0,0)	7842.734
(3,0,1)	7832.058
(3,0,0)	7843.357
(3,0,2)	7834.781
(2,0,1)	7829.578
(1,0,1)	7830.052
(2,0,0)	7841.194
(3,0,1)	7830.843
(2,0,2)	7829.616
(1,0,0)	7840.931
(1,0,2)	7829.695
(3,0,0)	7841.845
(3,0,2)	7833.578

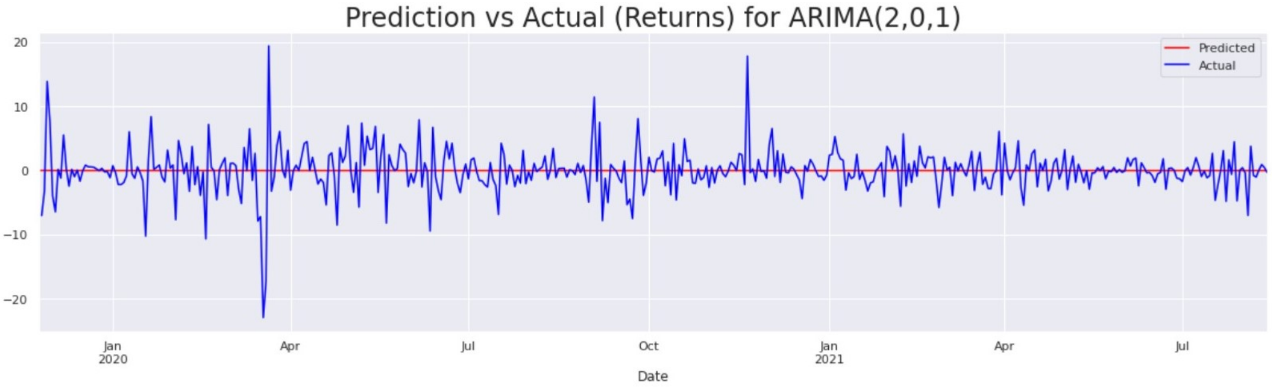
And the p-values and the corresponding estimates of different lags are given by:

Lag	Estimates	p-value
AR Lag 1	0.8516	0.000
AR Lag 2	-0.0398	0.055
MA Lag 1	-0.8672	0.000

We can see from the output tables ARIMA(2,0,1) has the least AIC, and also the p-values of corresponding estimates are also low, so we move forward with ARIMA(2,0,1) model.
So, the ARIMA(2,0,1) model is given by:

$$\hat{y_t} = 0.8516y_{t-1} - 0.0398y_{t-2} + \epsilon_t - 0.8672\epsilon_{t-1} \qquad ,t = 1, 2, \cdots, n$$

Also note that ARIMA(p,d,q) $\implies$ ARMA(p+d,q) , so the process reduces to ARMA(2,1).



10 Introduction: ARCH-GARCH

Stock prices prediction often emphasises on developing a model to predict mean behaviour of the data.But to achieve best results the stock market return should be characterised as a combination of drift and volatility. As risk uncertainty considerations of stock price pose a major concern, a new type of time series modelling technique has emerged to take account of conditional variances along with the mean behaviour. Engle introduced Autoregressive Conditional Heteroscedasticity model(ARCH) ,which provides us with a framework to model the time varying variances.

In financial time series it has been observed that in volatility (time varying variance) clustering large changes tend to follow large changes and small changes tend to follow small changes.This phenomenon is called conditional heteroscedasticity and can be modelled by ARCH model proposed by Engles(1982) a later generalization of it GARCH(Generalized ARCH)model proposed by Bollerslev(1986). While conventional time series econometric models drive under the assumption of constant mean, the ARCH process introduced allows the conditional variance to change over time as a function of past error leaving the unconditional variance constant. Several studies have further pointed out that the Autoregressive Integrated Moving Average (ARIMA) with ARCH errors are found to be successful in modelling some macroeconomic time series. In this section we will briefly discuss ARCH GARCH model and try to fit the same to our data.

10.1 Definition: ARCH model

The ARCH-type model is a non- linear model which we will try to define in terms of the distributions of the errors of a dynamic linear regression model. Suppose that we are modelling the variance of a series y_t . y_t can be modelled as

$$Y_t = x_t'\zeta + \epsilon_t, t = 1, 2, .., T$$

Where x_t is a $k \times 1$ vector of exogenous variables which may or may not include lagged values of dependent variable ζ is a $k \times 1$ vector of regression parameters. The distribution of stochastic error ϵ_t is conditioned on the realized values of set of variables

$$\Psi_{t-1} = \{y_{t-1}, x_{t-1}, y_{t-2}, x_{t-2}, ...\}$$

Engle assumed,

$$\epsilon_t|\Psi_{t-1} \sim N(0, h_t)$$

where,

$$h_t = \alpha_0 + \alpha_1\epsilon_{t-1}^2 + ... + \alpha_q\epsilon_{t-q}^2, \alpha_0 > 0 \text{ and } \alpha_i \geq 0, i = 1(1)q$$

so that the conditional variance is positive since,

$$\epsilon_{t-i} = y_{t-i} - x_{t-i}'\zeta, i = 1(1)q$$

, h_t clearly is a function of Ψ_{t-1}.

In ARCH regression model the variance of error ϵ_t is conditions on the realised value of lagged errors ϵ_{t-i} , i = 1,2,. . . ,q. Since variance is expected squared deviation, a linear combination of lagged squares is a natural measure of the recent trend in variance to translate to the current conditional variance h_t. The current error ϵ_t is an increasing function of magnitude of the lagged errors irrespective of their signs. So large errors of either sign tend to be followed by a large error of either sign and similarly small error of either sign tend to be followed by a small error of either sign. From the order of lag q the length of time for which shock persists in conditioning the variance of subsequent errors is determined. Larger value of q indicates longer episodes of volatility.

10.2 Definition: GARCH model

Bollerslev proposed an extension of conditional variance and suggested the conditional variance to be

$$h_t = \alpha_0 + \alpha_1\epsilon_{t-1}^2 + ... + \alpha_q\epsilon_{t-q}^2 + \beta_1 h_{t-1} + ... + \beta_p h_{t-p} \qquad (1)$$

where,

$$\alpha_0 > 0 \text{ and } \alpha_i \geq 0, i = 1(1)q$$
$$\beta_i \geq 0, i = 1(1)p,$$

so that the conditional variance is strictly positive. A GARCH process with orders p and q is denoted as GARCH(p,q). Now we write h_t as

$$h_t = \alpha_0 + \alpha(B)\epsilon_t^2 + \beta(B)h_t$$

where ,

$$\alpha(B) = \alpha_1 B + .. + \alpha_q B^q$$
$$\beta(B) = \beta_1 B + .. + \beta_p B^p$$

are polynomials in the backshift operator B. If the roots of $1 - \beta(Z)$ lies outside the unit circle we can rewrite (1) as

$$h_t = \frac{\alpha_0}{1 - \beta(1)} + \frac{\alpha(B)}{1 - \beta(B)}\epsilon_t^2$$
$$= \alpha_0^* + \sum_{i=1}^{\infty} \delta_i \epsilon_{t-i}^2 \qquad (2)$$

where $\alpha_0^* = \frac{\alpha_0}{1-\beta(1)}$, δ_i is the coefficients of B^i in the expansion of $\frac{\alpha(B)}{1-\beta(B)}$. From the expression (2) we can see that a GARCH(p,q) process is an infinite order ARCH process. GARCH can parsimoniously represent a high order ARCH process.

11 ARIMA-ARCH Model

The ARIMA-ARCH model is one model in which the variance of the error term of the ARIMA model follows an ARCH process. In our data we have already fitted ARIMA(2,0,1) model. The model is given by ,

$$\hat{y}_t = 0.8516y_{t-1} - 0.0398y_{t-2} + \epsilon_t - 0.8672\epsilon_{t-1} \qquad\qquad , t = 1, 2, .., n$$

Here y_t are the observed series and ϵ_t are the model residuals. Now we consider the squared residuals after fitting ARIMA(2,0,1) to our data , i.e we consider $r_t^2 = (y_t - \hat{y}_t)^2$. for $t = 1, 2, ..., n$. We will apply Engle's Lagrange Multiplier Test to test for the existence of ARCH-Effect in the series r_t^2.

11.1 ARCH effect in ARIMA residuals

The detection of the ARCH effect in a time series is a test of serial independence applied to the serially uncorrelated fitting error of some model,in our case ARIMA model. We have assumed that linear serial dependence inside the original series is removed with an efficient model. Hence, any further serial dependence must be due to some nonlinear mechanism which has not been detected by the model. Here, the nonlinear mechanism we are concerned with is the conditional heteroskedasticity.

- **Lagrange Multiplier Test**

In this testing process,

$$H_0 : \text{ARCH-Effect is not present}$$
$$H_1 : \text{ARCH-Effect is present}$$

This procedure simply involves obtaining the squares of the residual from fitted model r_t^2 and regress them on a constant and p lagged values, where is the ARCH lags. Let us consider the equation:

$$r_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i r_{t-i}^2$$

The hypothesis is that, in the absence of ARCH components, we have $\alpha_i = 0$ for all i=1, 2, . . . ,p, against the alternative that, in the presence of ARCH components, at least one of the estimated α_i must be significant.
The test statistic is given by nR^2. Here R is the sample multiple correlation coefficient computed from the above regression of r_t^2 on $r_{t-1}^2, ..., r_{t-p}^2$.
Under the null hypothesis H_0 (ARCH effect is not present), the nR^2 asymptotically follows a χ^2 distribution with degrees of freedom p.

- **Conclusion:**

We tested the hypothesis at level of significance 0.05 and obtain the p_value as $3.45e - 69 < 0.05$. Based on the given observation we reject the null hypothesis at 5% level of significance and conclude on that the squared residual of ARIMA(2,0,2) model contains ARCH-Effect. Now we proceed further to find the order of ARCH model to the squared residuals.

As the ARCH model can capture the behavior of conditional variance rather than mean behavior, it can be used in our case to deal with the ARCH effect in the residual series.

11.2 Model Building

Now it is important to select an appropriate ARCH model. We have summarized the algorithm below. Later we will discuss how an appropriate model can be constructed.

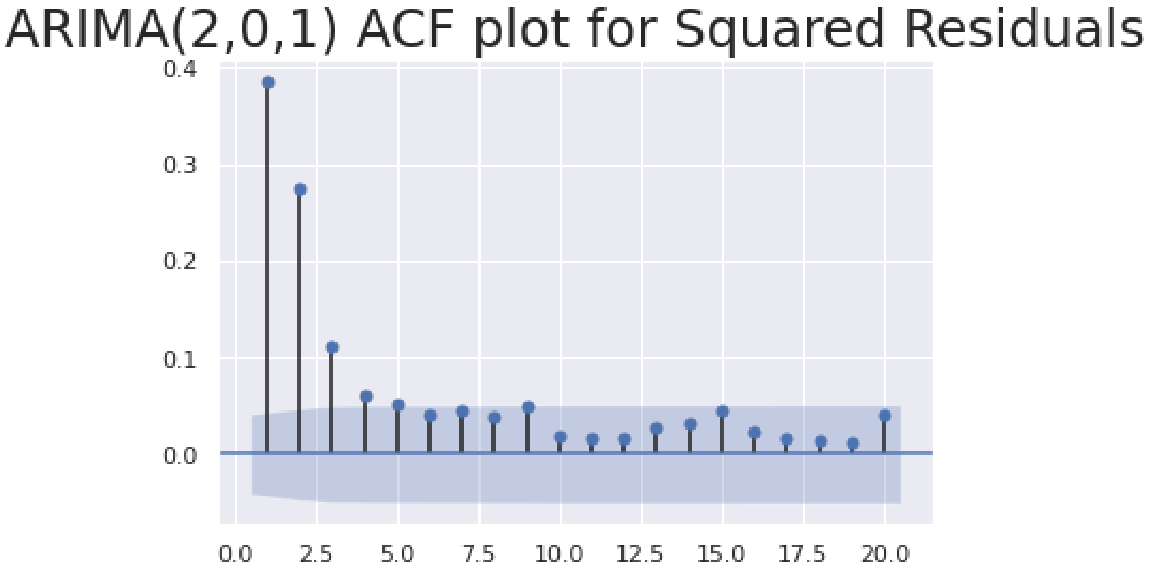
11.2.1 Algorithm

The algorithm for ARIMA-ARCH model will be given by :

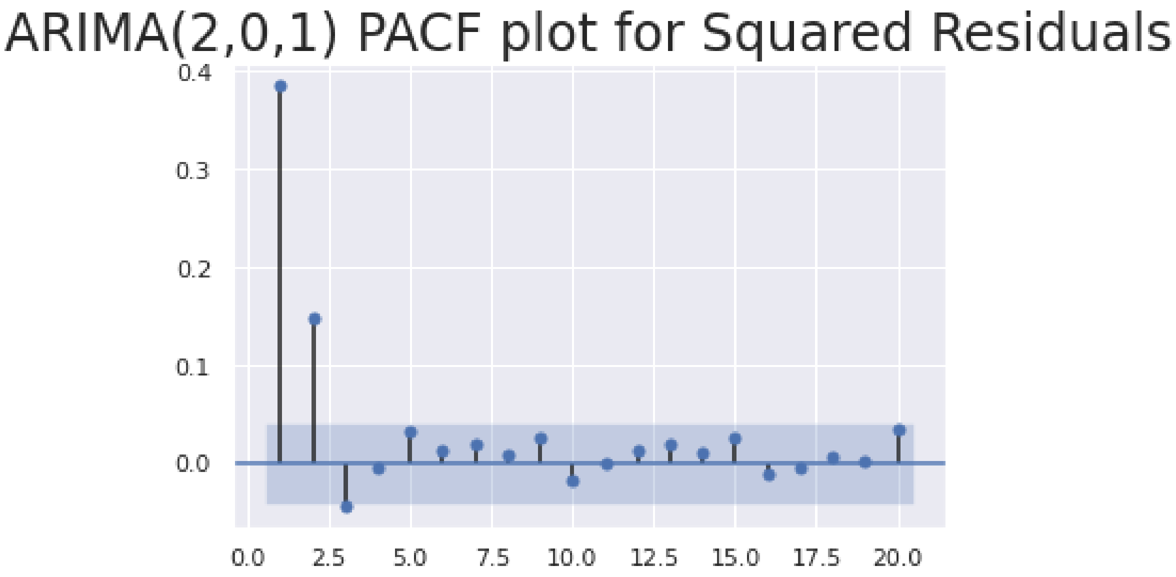
- Fit an appropriate ARIMA model to the stationary data.In this case we have fitted ARIMA(2,0,1) model to our data.
- Obtain the residuals after fitting the ARIMA model. From this series of residuals we need to obtain squared residuals.
- Fit an appropriate ARCH model to the residual of ARIMA model.

11.2.2 Order Selection

At first we examine the ACF plot of the squared residuals.



The ACF plot indicates significant lag up to order 4 or 5. We will proceed to check the PACF plot for the same.



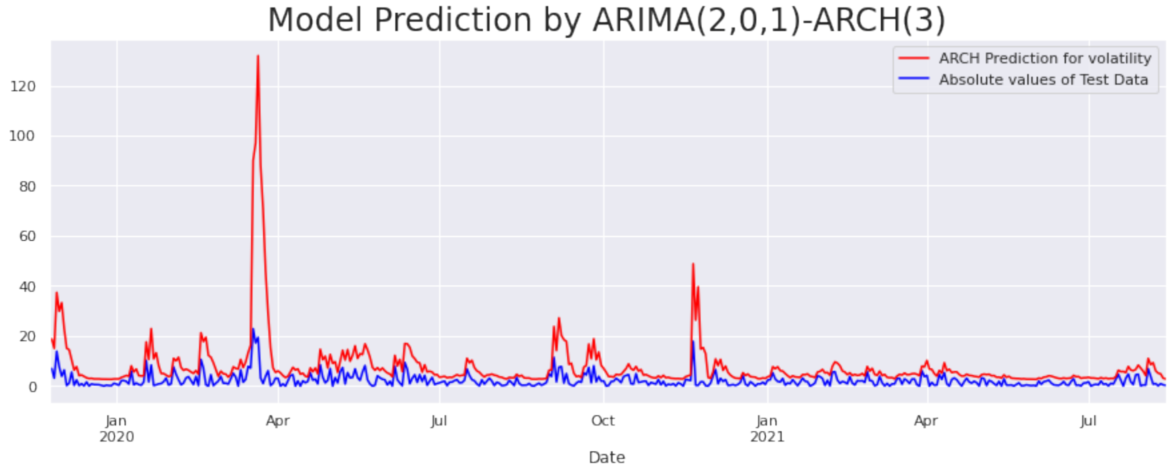
The PACF plot also indicates spikes up to 2 or 3. Based on the observations we will fit ARCH(6) model and examine the significance of the coefficient for each lag.

Lag	Estimate	Standard Error	p_value
1	0.1356	3.912e-02	5.292e-04
2	0.0709	3.539e-02	4.515e-02
3	0.1095	4.309e-02	1.103e-02
4	0.0347	3.757e-02	0.356
5	0.0347	2.871e-02	0.226
6	0.0310	3.396e-02	0.361

The table suggests that up to lag 3 the estimates of the coefficients are significant as the p_value associated with them are less than 0.05. On the other hand the for lags greater than 3 the coefficients are not significant. Hence we resort to fit the model of order 3 to the residuals of ARIMA(2,0,1). The ARCH(3) model is given by ,

$$\hat{v}_t = 2.9940 + 0.1356\epsilon^2_{t-1} + 0.0709\epsilon^2_{t-2} + 0.1095\epsilon^2_{t-3} \qquad ,t = 1,2,..,n$$

11.3 Volatility Prediction by ARCH(3)



From the figure it can be understood that the fitted model has captured the model volatility.

12 ARIMA-GARCH Model

The ARIMA-GARCH model is one model in which the variance of the error term of the ARIMA model follows an GARCH process. In our data we have already fitted an ARIMA(2,0,1) model. Proceeding in a similar way as before we do the model building.

12.1 Model Buliding

To select an appropriate GARCH model we follow the following algorithm.

12.1.1 Algorithm and Order Selection

- With the residuals of ARIMA(2,0,1) model we fit GARCH(1,1) model

Parameters	Coefficients	p-value
α_1	0.1109	2.531e-03
β_1	0.7722	3.817e-23

  As α_1 and β_1 are both significant, we try to fit GARCH(1,2) model

- For GARCH(1,2) model we find the following:

	Coefficients	p-value
α_1	0.1394	5.346e-04
β_1	0.2255	0.590
β_2	0.5074	0.260

  As β_2 is insignificant, we fix q to be 1 and proceed to fit GARCH(2,1)

- For GARCH(2,1) model we find the following:

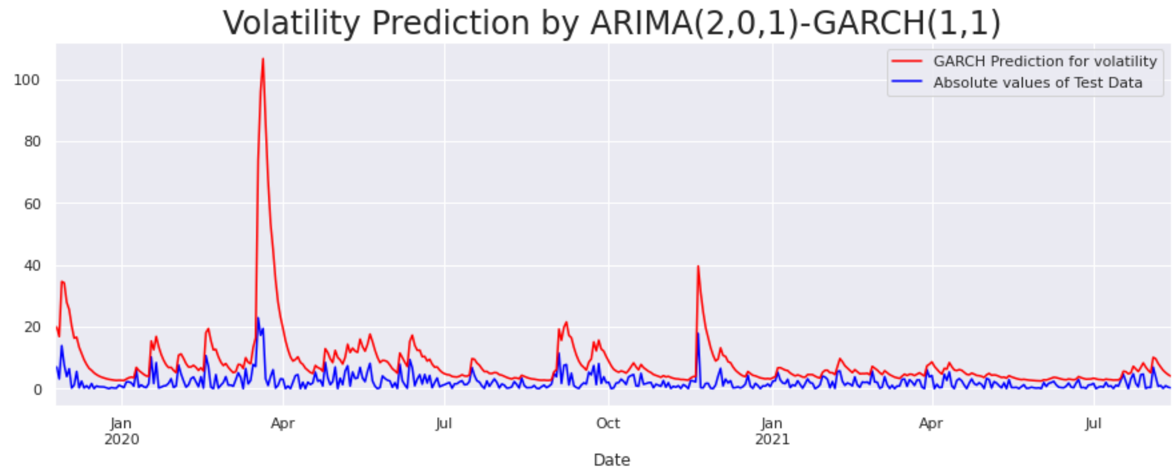
	Coefficients	p-value
α_1	0.1109	2.466e-03
α_2	6.9086e-14	1
β_1	0.7722	1.118e-09

  As α_2 is insignificant, we fix p to be 1 and stick to GARCH(1,1) model.

Hence the fitted GARCH(1,1) model is given by,

$$\hat{h}_t = 0.1109\epsilon_{t-1}^2 + 0.7722h_{t-1}$$

12.2 Volatility Prediction by GARCH(1,1)



From the figure it can be understood that the fitted model has captured the model volatility.

13 Conclusion

In this work our central purpose was to predict the closing stock price of the chosen company. Using appropriate methods we have successfully transformed the non-stationary raw data into a stationary one. After that, under the assumption of constant error variance we have identified that the AR(7), MA(7) and ARIMA(2,0,1) models are best to describe the mean nature of the stock price. Later, we detected the presence of ARCH-effect in the ARIMA(2,0,1) residuals through Engle's Lagrange Multiplier test. To capture the volatility along with the mean we have used the combination of ARIMA and ARCH(GARCH) and fitted Hybrid ARIMA(2,0,1)-ARCH(3) and ARIMA(2,0,1)-GARCH(1,1) to the data.

Reference:

1. Lecture notes on Time Series Analysis(MTH517A) by Prof. Amit Mitra.

2. "*Testing and modelling autoregressive conditional heteroskedasticity of streamflow processes*", Nonlinear Processes in Geophysics (2005) 12: 55–66: W.Wang, P.H.A.J.M. Van Gelder, J.K. Vrijling, J.Ma

3. "*Hybrid of ARIMA-ARCH Modelling of Daily Share Price Data of Okomuc Oil Plc in NigeriaHybrid of ARIMA-ARCH Modelling of Daily Share Price Data of Okomu Oil Plc in Nigeria*", Journal of the Nigerian Association of Mathematical Physics Volume36, No. 2 (July, 2016), pp163 – 168 : Osemwenkhae J.E., Eguasa E.B. and Iduseri A.

4. Youtube Playlist - *Time Series Analysis in Python, Data Ranger*

5. Google