NAME : Hrutuja Patkar
SEAT NO. : 2115205

SUBJECT : TIME SERIES
SUBJECT INCHARGE : MR. ASHISH MHATRE

DATE OF SUBMISSION : 14 FEBRUARY 2022

Q.1 Estimate parameters of AR (1) model by MLE method.

Consider the stationary AR(1) model

$$\begin{aligned} y_t &= c + \phi y_{t-1} + \varepsilon_t, \;\; \varepsilon_t \sim iid\ N(0, \sigma^2), \;\; t = 1, \ldots, T \\ \boldsymbol{\theta} &= (c, \phi, \sigma^2)', \;\; |\phi| < 1 \end{aligned}$$

Conditional on I_{t-1}

$$y_t | I_{t-1} \sim N(c + \phi y_{t-1}, \sigma^2), \;\; t = 2, \ldots, T$$

which only depends on y_{t-1}. The conditional density $f(y_t|I_{t-1}, \theta)$ is then

$$f(y_t|y_{t-1}, \boldsymbol{\theta}) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(y_t - c - \phi y_{t-1})^2\right), \;\; t = 2, \ldots, T$$

To determine the marginal density for the initial value y_1, recall that for a stationary AR(1) process

$$\begin{aligned} E[y_1] &= \mu = \frac{c}{1-\phi} \\ var(y_1) &= \frac{\sigma^2}{1-\phi^2} \end{aligned}$$

It follows that

$$\begin{aligned} y_1 &\sim N\left(\frac{c}{1-\phi}, \frac{\sigma^2}{1-\phi^2}\right) \\ f(y_1; \boldsymbol{\theta}) &= \left(2\pi\frac{\sigma^2}{1-\phi^2}\right)^{-1/2} \exp\left(-\frac{1-\phi^2}{2\sigma^2}\left(y_1 - \frac{c}{1-\phi}\right)^2\right) \end{aligned}$$

The conditional log-likelihood function is

$$\begin{aligned} \sum_{t=2}^{T} \ln f(y_t|y_{t-1}, \boldsymbol{\theta}) &= \frac{-(T-1)}{2}\ln(2\pi) - \frac{(T-1)}{2}\ln(\sigma^2) \\ &\quad -\frac{1}{2\sigma^2}\sum_{t=2}^{T}(y_t - c - \phi y_{t-1})^2 \end{aligned}$$

Notice that the conditional log-likelihood function has the form of the log-likelihood function for a linear regression model with normal errors. It follows that the conditional mles for c and ϕ are identical to the least squares estimates from the regression

$$y_t = c + \phi y_{t-1} + \varepsilon_t, t = 2, \ldots, T$$

and the conditional mle for σ^2 is

$$\hat{\sigma}^2_{cmle} = (T-1)^{-1} \sum_{t=2}^{T} (y_t - \hat{c}_{cmle} - \hat{\phi}_{cmle} y_{t-1})^2$$

The marginal log-likelihood for the initial value y_1 is

$$\ln f(y_1; \boldsymbol{\theta}) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\frac{\sigma^2}{1-\phi^2}\right) - \frac{1-\phi^2}{2\sigma^2}\left(y_1 - \frac{c}{1-\phi}\right)^2$$

The exact log-likelihood function is then

$$\begin{aligned}\ln L(\boldsymbol{\theta}|\mathbf{y}) &= -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\frac{\sigma^2}{1-\phi^2}\right) - \frac{1-\phi^2}{2\sigma^2}\left(y_1 - \frac{c}{1-\phi}\right)^2 \\ &\quad -\frac{(T-1)}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=2}^{T}(y_t - c - \phi y_{t-1})^2\end{aligned}$$

The exact log-likelihood function is a non-linear function of the parameters θ, and so there is no closed form solution for the exact mles. The exact mles must be determined by numerically maximizing the exact log-likelihood function. Usually, a Newton-Raphson type algorithm is used for the maximization which leads to the interative scheme

$$\hat{\boldsymbol{\theta}}_{mle,n} = \hat{\boldsymbol{\theta}}_{mle,n-1} - \hat{\mathbf{H}}(\hat{\boldsymbol{\theta}}_{mle,n-1})^{-1}\hat{\mathbf{s}}(\hat{\boldsymbol{\theta}}_{mle,n-1})$$

where $\hat{\mathbf{H}}(\hat{\boldsymbol{\theta}})$ is an estimate of the Hessian matrix (2nd derivative of the log-likelihood function), and $\hat{\mathbf{s}}(\hat{\boldsymbol{\theta}})$ is an estimate of the score vector (1st derivative of the log-likelihood function). The estimates of the Hessian and Score may be computed numerically (using numerical derivative routines) or they may be computed analytically (if analytic derivatives are known).

Q.2 Select any real-life data in which both trend and seasonal components are present. Apply a smoothing technique to smooth the chosen data using appropriate smoothing constants.

TOPIC: To analyse the consumption of coal over a period of time.

OBJECTIVE : To smoothen the data in order to find clear trends to analyse the consumption of coal for years(2014-2020).

WHY THIS TOPIC ?

Coal is a fossil fuel that poses serious threats to our environment. Though it has the lowest consumption rate among all fossil fuels, it is necessary to understand how we have been consuming it over years.

Understanding the consumption of coal over time, can help us determine if coal is really causing the threat to our environment, or we have reduced its consumption for a better alternative and also to safeguard our planet.

DATA SOURCE :

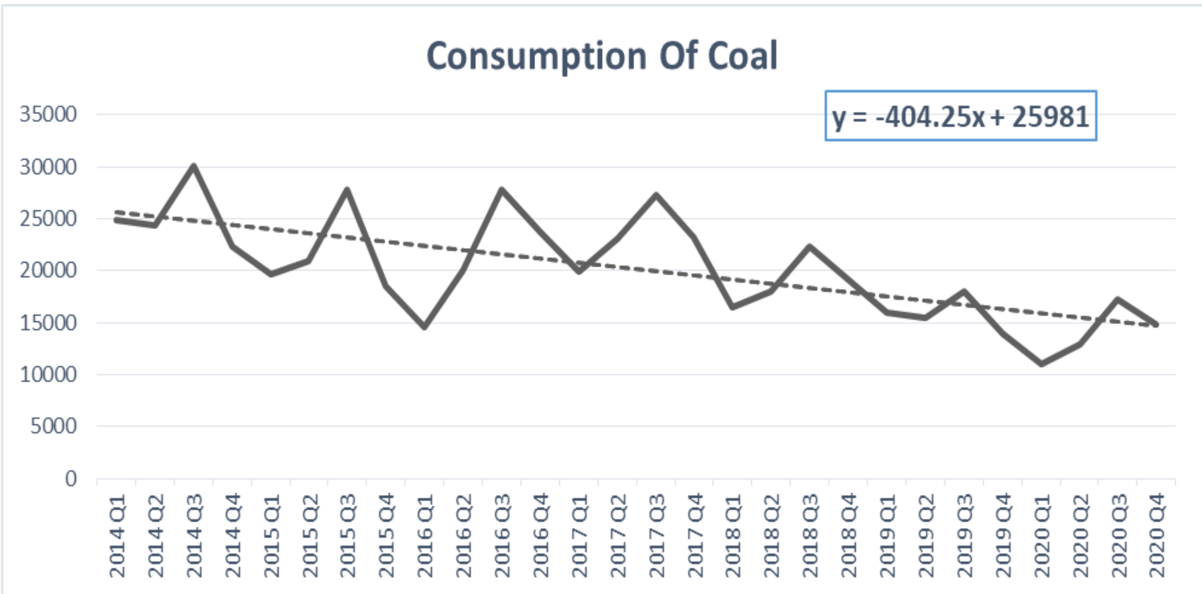
https://github.com/rishabh89007/Time_Series_Datasets

Total consumption of coal for quarterly (short tons) over years:

Year	Q1	Q2	Q3	Q4
2014	24914.333	24410.19	30060.366	22272.614
2015	19639.959	20880.065	27792.799	18466.374
2016	14616.943	20077.346	27738.163	23698.04
2017	19887.551	23138.725	27312.98	23148.298
2018	16466.977	17975.559	22259.525	19196.961
2019	16002.251	15464.589	17977.253	13866.538
2020	11029.285	12950.538	17219.614	14869.537

INTERPRETATION :

Plot for the consumption of coal over years 2014-2020:



The graph shows that there is a decreasing trend and seasonality present in the time series.

The Holts and Winters exponential smoothing is used to when there is trend as well as seasonality present in the time series.

ANALYSIS :

Step 1: The trend line is given by equation: **y = -404.25x + 25981**

Year	Q1	Q2	Q3	Q4	Mean	Estimated y
2014	24914.333	24410.19	30060.366	22272.614	25414.37575	-10247780.4
2015	19639.959	20880.065	27792.799	18466.374	21694.79925	-8744141.597
2016	14616.943	20077.346	27738.163	23698.04	21532.623	-8678581.848
2017	19887.551	23138.725	27312.98	23148.298	23371.8885	-9422104.926
2018	16466.977	17975.559	22259.525	19196.961	18974.7555	-7644563.911
2019	16002.251	15464.589	17977.253	13866.538	15827.65775	-6372349.645
2020	11029.285	12950.538	17219.614	14869.537	14017.2435	-5640489.685

Step2:
Estimation of trends:

Year	II - b/4	yt - b/8	yt + b/8	III + b/4
2014	-10257523.3	-10251028	-10244533	-10238037.52
2015	-8753884.47	-8747389.22	-8740894	-8734398.722
2016	-8688324.72	-8681829.47	-8675334.2	-8668838.973
2017	-9431847.8	-9425352.55	-9418857.3	-9412362.051
2018	-7654306.79	-7647811.54	-7641316.3	-7634821.036
2019	-6382092.52	-6375597.27	-6369102	-6362606.77
2020	-5650232.56	-5643737.31	-5637242.1	-5630746.81

Step 3:
Detrend the data:

The multiplicative method is preferred when the seasonal variations are changing over time.For multiplicative case, we obtain seasonal indices using the equation given by, $S_i = \bar{X}_i / \bar{X}$

Calculation = Observed value / Trend estimated value						
Trend Eliminated Value						
Year	I	II	III	IV		
2014	-0.00242888	-0.00238124	-0.0029343	-0.002175477		
2015	-0.00224357	-0.00238701	-0.0031796	-0.002114212		
2016	-0.00168237	-0.00231257	-0.0031974	-0.002733704		
2017	-0.00210855	-0.00245495	-0.0028998	-0.002459351		
2018	-0.00215133	-0.00235042	-0.002913	-0.002514396		
2019	-0.00250737	-0.00242559	-0.0028226	-0.00217938		
2020	-0.00195201	-0.00229467	-0.0030546	-0.002640775		
Mean Xi	-0.00215344	-0.00237235	-0.0030002	-0.002402471	Mean X =	-0.002482113
Si = Xi/X	0.86758361	0.95577844	1.2087243	0.967913652	Sum Si =	4

Step 4:

Constants:

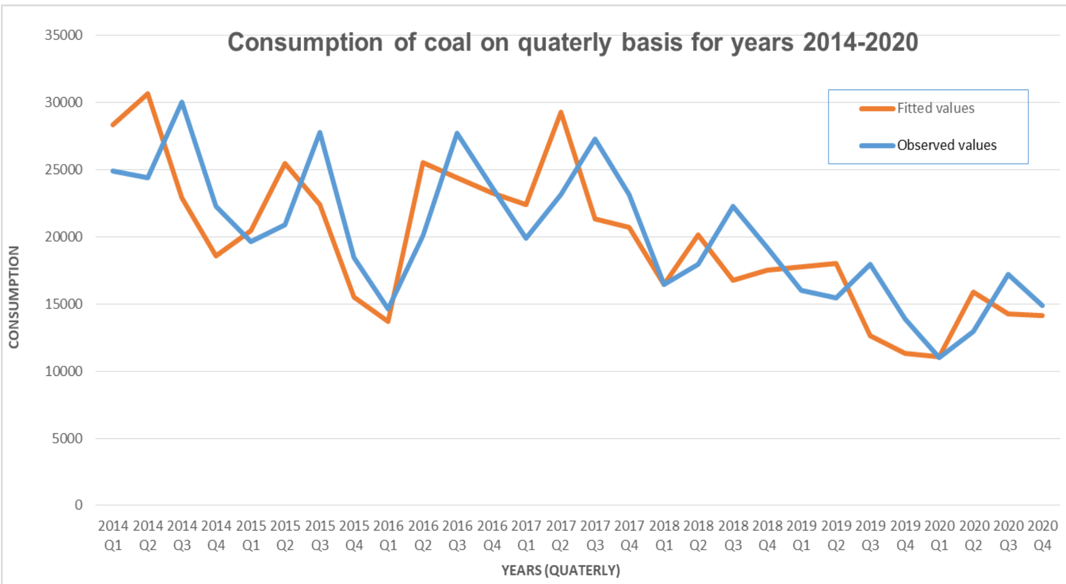
alpha	0.79014062
beta	0.80080494
gamma	0.01
RMSE	3797.28555

Year	T	Observation	Lt	Tt	St	Ycap t+p	Error
0	-3				0.86758361		
	-2				0.95577844		
	-1				1.2087243		
	0		25981	-404.25	0.96791365		
2014	1	24914.333	28057.9294	1582.690712	0.867787379	28329.86564	-3415.53
	2	24410.19	26400.2301	-1012.229628	0.955466859	30687.09307	-6276.9
	3	30060.366	24978.3102	-1340.311573	1.208671644	22879.54157	7180.824

	4	22272.614	23142.5432	-1737.074768	0.967858613	18575.39533	3697.219
2015	1	19639.959	22374.7784	-960.846525	0.867887227	20460.30225	-820.343
	2	20880.065	21761.0623	-682.8627848	0.95550734	25476.62207	-4596.56
	3	27792.799	22592.3456	529.6728964	1.208886793	22378.84477	5413.954
	4	18466.374	19927.9548	-2028.149096	0.967446595	15535.01276	2931.361
2016	1	14616.943	17063.978	-2697.484064	0.86777432	13727.29041	889.6526
	2	20077.346	19617.5649	1507.599492	0.956186639	25537.93221	-5460.59
	3	27738.163	22563.2574	2659.231519	1.209091435	24401.41104	3336.752
	4	23698.04	24648.0262	2199.198915	0.967386708	23297.33251	400.7075
2017	1	19887.551	23742.4929	-287.0859598	0.86747293	22427.74676	-2540.2
	2	23138.725	24042.9207	183.3979423	0.956248697	29291.83439	-6153.11
	3	27312.98	22933.1384	-852.1872309	1.208910352	21360.81861	5952.161
	4	23148.298	23540.9258	316.9676781	0.967546056	20696.07672	2452.221
2018	1	16466.977	20005.8038	-2767.804711	0.8670293	16483.8142	-16.8372
	2	17975.559	18470.6159	-1780.719135	0.956418186	20176.58893	-2201.03
	3	22259.525	18051.2982	-690.5020736	1.209152508	16797.36986	5462.155

	4	19196.961	19320.4077	878.7645322	0.9678067	17513.27418	1683.687
2019	1	16002.251	18822.1475	-223.9636752	0.866860827	17787.64126	-1785.39
	2	15464.589	16679.0046	-1760.851902	0.956125894	18038.32175	-2573.73
	3	17977.253	14878.2463	-1792.809126	1.209143894	12664.17379	5313.079
	4	13866.538	14067.0762	-1006.707769	0.967986075	11321.52177	2545.016
2020	1	11029.285	12793.9956	-1220.020418	0.866812892	11066.17738	-36.8924
	2	12950.538	13131.2072	27.01867642	0.956427047	15910.18851	-2959.65
	3	17219.614	14013.8976	712.2448374	1.209339982	14254.70086	2964.913
	4	14869.537	15228.0164	1114.147986	0.968070806	14165.59881	703.9382

OUTPUT:



CONCLUSION:

Thus, the time series now shows a clear decreasing trend even after removing the trend and seasonality components.

This draws a conclusion that the consumption of coal has reduced over a time period and thus we can say we may not have much threat to the environment from consumption of coal.

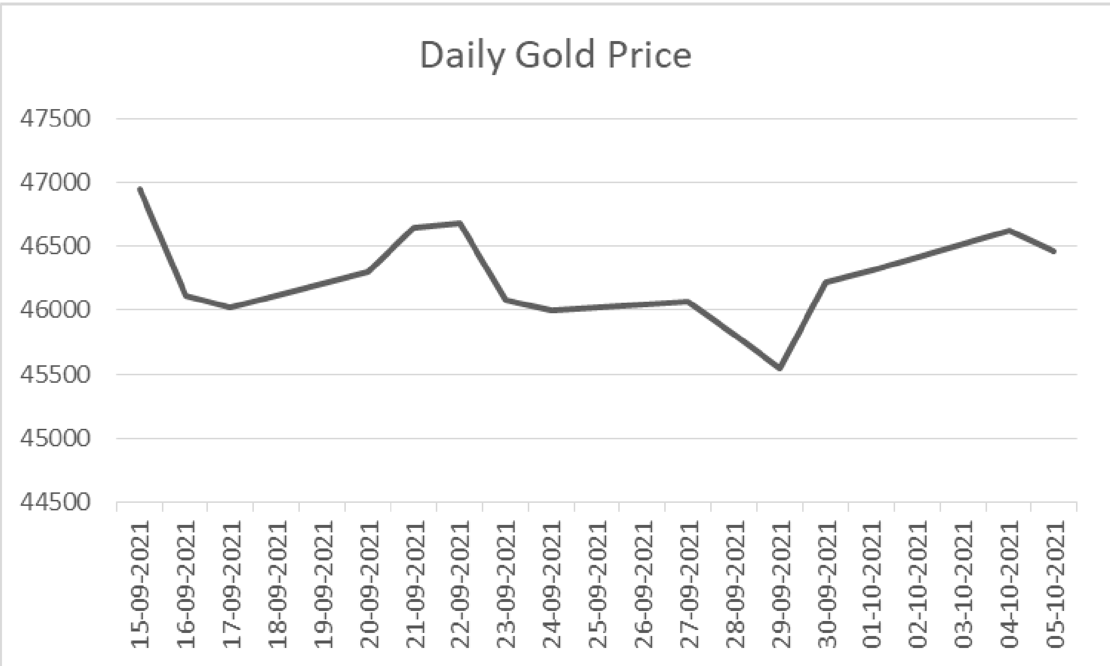
Q.3 Select any real-life time series data and for chosen data obtain acf, pacf. Plot the graphs. If the chosen time series is stationary, then identify which model can be fitted to the data. If it is non-stationary, then convert it into stationary and then identify which model can be fitted to the data.

TOPIC: Daily gold prices

OBJECTIVE: To find if the data is stationary, and if not ,convert it into stationary and fit a model to it by using acf and pacf plots.

DATA SOURCE :
https://www.kaggle.com/datasets?search=gold+prices

INTERPRETATION & ANALYSIS :



From the graph, we can see that the mean,variance and covariance is not constant over time. Hence the time series is **non-stationary.**

Also, by using **Dickey-Fuller Test:**

> adf.test(Price)

Augmented Dickey-Fuller Test

data: Price
Dickey-Fuller = -1.6348, Lag order = 2, p-value = 0.7115
alternative hypothesis: stationary

From the p-value, we can say that the time series is non-stationary.

Converting Non-stationary data into stationary using Transformations and Differencing methods:

Using 1st, 2nd and 3rd order differencing:

ddprice = diff(Price)
> adf.test(ddprice)

Augmented Dickey-Fuller Test

data: ddprice
Dickey-Fuller = -2.5835, Lag order = 2, p-value = 0.3501
alternative hypothesis: stationary

> dd2price = diff(Price, differences = 2)
> adf.test(dd2price)

Augmented Dickey-Fuller Test

data: dd2price
Dickey-Fuller = -3.0793, Lag order = 2, p-value = 0.1612
alternative hypothesis: stationary

> dd3price = diff(Price, differences = 3)
> adf.test(dd3price)

Augmented Dickey-Fuller Test

data: dd3price
Dickey-Fuller = -2.1951, Lag order = 2, p-value = 0.4981
alternative hypothesis: stationary
By looking at the p-values of all the order differencing, we find none of them to be stationary.

Using log-transformation:

> lnprice = log(Price)
> adf.test(lnprice)

Augmented Dickey-Fuller Test

data: lnprice
Dickey-Fuller = -1.6314, Lag order = 2, p-value = 0.7128
alternative hypothesis: stationary

By looking at the p-value of the log-transformation series, we find data to be non-stationary.

Using 1st, 2nd and 3rd order differencing on log transformation:

> dprice = diff(lnprice)
> adf.test(dprice)

Augmented Dickey-Fuller Test

data: dprice
Dickey-Fuller = -2.5758, Lag order = 2, p-value = 0.353
alternative hypothesis: stationary

> d2price = diff(lnprice, differences = 2)
> adf.test(d2price)

Augmented Dickey-Fuller Test

data: d2price
Dickey-Fuller = -3.0812, Lag order = 2, p-value = 0.1605
alternative hypothesis: stationary

> d2price = diff(lnprice, differences = 3)
> d3price = diff(lnprice, differences = 3)
> adf.test(d3price)

Augmented Dickey-Fuller Test

data: d3price
Dickey-Fuller = -2.2087, Lag order = 2, p-value = 0.4929
alternative hypothesis: stationary
By looking at the p-values of all the order differencing, we find none of them to be stationary.

Using square root-transformation:

sqq = sqrt(Price)
> adf.test(sqq)

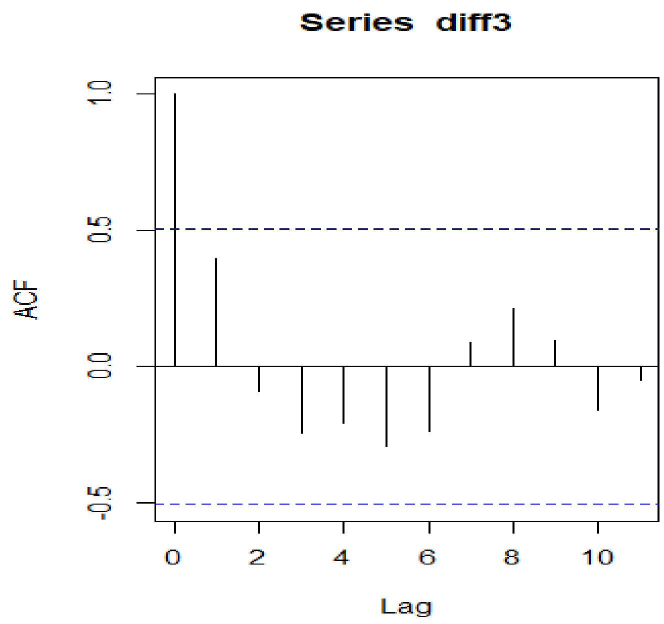
Augmented Dickey-Fuller Test

data: sqq
Dickey-Fuller = -1.6331, Lag order = 2, p-value = 0.7121
alternative hypothesis: stationary

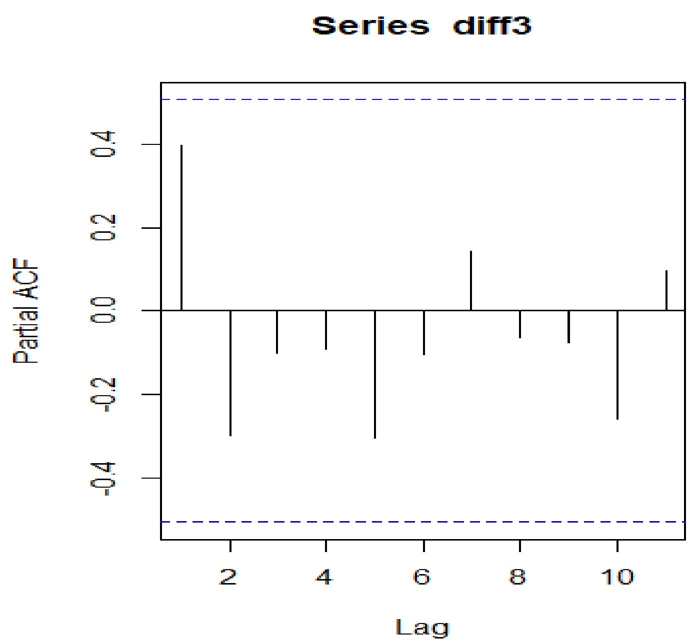
By looking at the p-value of the square root transformation series, we find data to be non-stationary.

Finding ACF and PACF :

ACF:



PACF:



CONCLUSION :

Even though we find ACF & PACF plots for time, we are not able to fit any model to the time series as it is non-stationary.
The given non-stationary time series data can be made stationary by trying various transformations and taking their order differentiation or solved by linear models.