Faculty of Technology, Design & Environment

Assessment cover

STUDENTS, PLEASE COPY THIS PAGE AND USE AS THE COVER PAGE FOR YOUR SUBMISSION

Module No:	DALT7010	Module title:	**Time Series Analysis**

Assessment title:	Coursework Sections 1 and 2

Due date and time**:**	3 May 2024 5PM

Estimated total time to be spent on assignment:	76 hours per student

LEARNING OUTCOMES

On successful completion of this module, students will be able to achieve the module following learning outcomes (LOs): *LO numbers and text copied and pasted from the module descriptor*
LO1 Demonstrate understanding of the dynamic nature of the interrelationships between deterministic trend, seasonality and stochasticity within time series models.
LO2 Define and devise autoregressive integrated moving average (ARIMA) and seasonal ARIMA (SARIMA) models and evaluate their properties.
LO3 Apply Box-Jenkins methodology to select appropriate models for time series data, critically evaluate the merits of outcomes and create solutions for shortcomings.

Engineering Council AHEP4 LOs assessed (from S1 2022-23) *LOs copied and pasted from the AHEP4 matrix*	
M1	Apply a comprehensive knowledge of mathematics, statistics, natural science and engineering principles to the solution of complex problems. Much of the knowledge will be at the forefront of the particular subject of study and informed by a critical awareness of new developments and the wider context of engineering
M2	Formulate and analyse complex problems to reach substantiated conclusions. This will involve evaluating available data using first principles of mathematics, statistics, natural science and engineering principles, and using engineering judgment to work with information that may be uncertain or incomplete, discussing the limitations of the techniques employed
M3	Select and apply appropriate computational and analytical techniques to model complex problems, discussing the limitations of the techniques employed
M17	Communicate effectively on complex engineering matters with technical and non-technical audiences, evaluating the effectiveness of the methods used

Statement of Compliance *(please tick to sign)*

☐ I declare that the work submitted is my own and that the work I submit is fully in accordance with the University regulations regarding assessments *(www.brookes.ac.uk/uniregulations/current)*

School of Engineering, Computing & Mathematics

A. Stationarity of time series

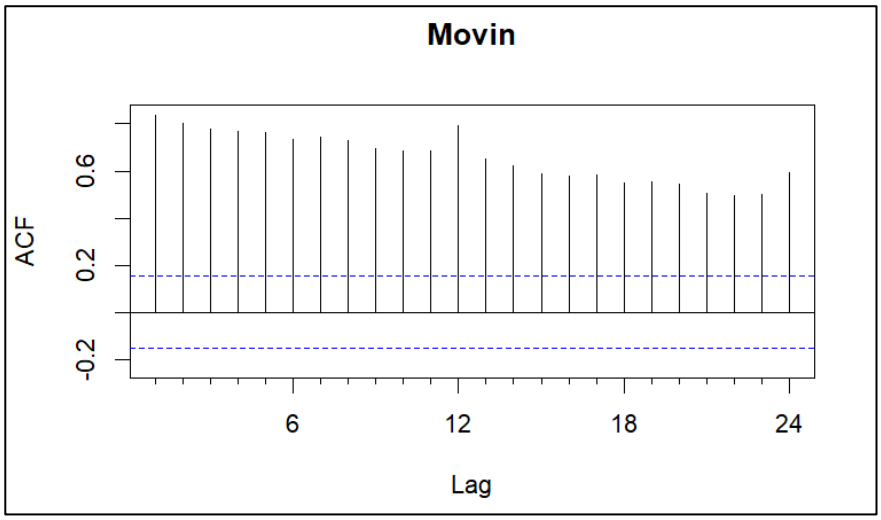

Fig 1: ACF plot of Movin timeseries

As per fig (timeseries) plot the trend seems increasing. Also, by looking at the autocorrelation function plot (ACF) of the Movin timeseries, the spikes decay slowly as the lags increases and there is not sudden decay after a certain number of lags which implies that the timeseries is non-stationary.

Also, by looking at the Movin timeseries by ignoring the seasonal spikes at regular intervals which shows seasonality, the other two characteristics are expanding variance and positive trend.

So, the series is neither mean nor variance stationary.

B. Explain possible differencing and transforming of the series to make it mean and variance stationary, if required.

i. The timeseries is transformed using log transformation.
ii. Then the timeseries is differenced once using lag = 1 to make it mean stationary.
iii. Box-Cox method is used to then transform the mean stationary timeseries to make it variance stationary.
iv. Ljung-Box test is carried out to check the variance stationarity.

```
> Box.test(residuals_squared, lag = 12, type = "Ljung-Box")

        Box-Ljung test

data:  residuals_squared
X-squared = 8.7099, df = 12, p-value = 0.7275
```

The higher p value of 0.7275 states that there is no significant evidence of autocorrelation. Thus, this Ljung-Box test implies that the variance does not change over time and is constant which indicates variance stationarity is achieved.

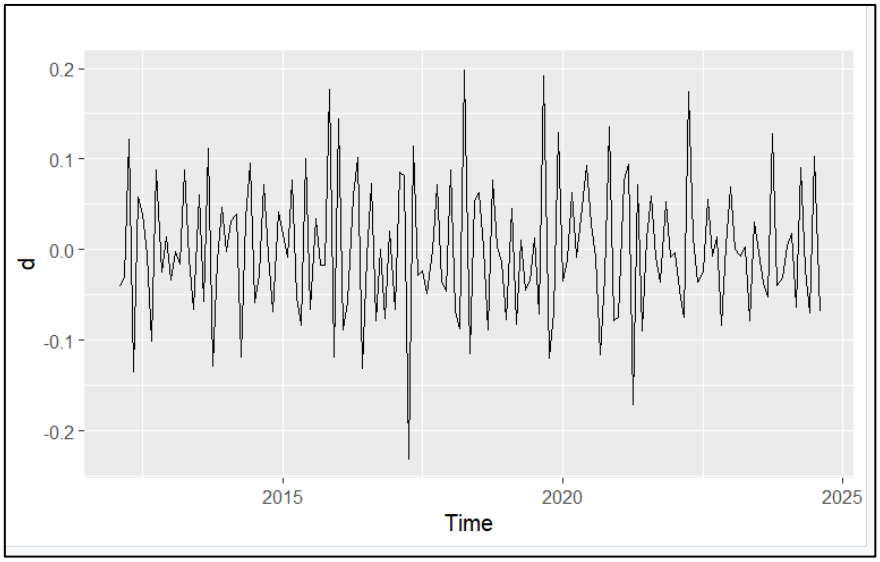

Fig 2: Transformed Movin timeseries

Now, after the ln transformation and one regular and only one seasonal differencing the timeseries has achieved stationarity i.e. mean and variance stationarity and is ready for forecasting.

C. Investigate the sample autocorrelation function and partial autocorrelation function to determine the order of SARIMA model. Clearly explain the stages you follow.

Now that the is model ready for further investigation, lets consider the ACF and PAC plots to determine the order of the model.

STAGE 1. Identification:

To find out the appropriate value of regular structure p and seasonal order P and Q using the ACF and PACF plots.

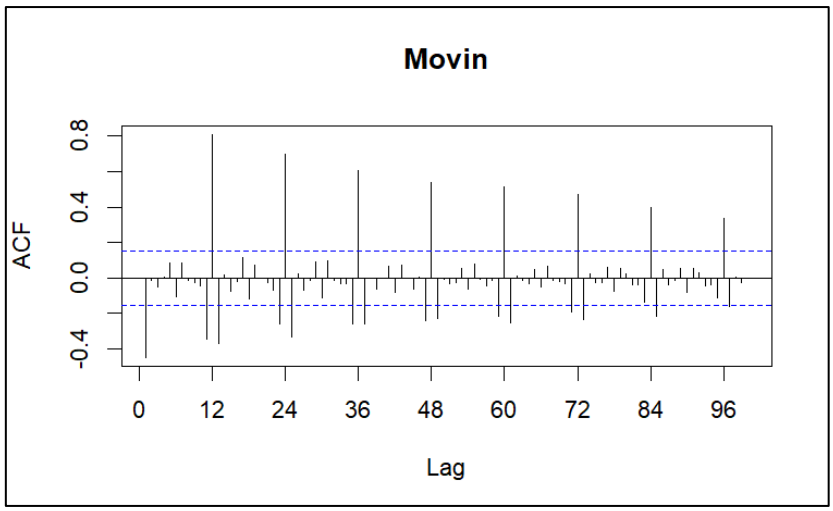

Fig 3: ACF plot of transformed timeseries

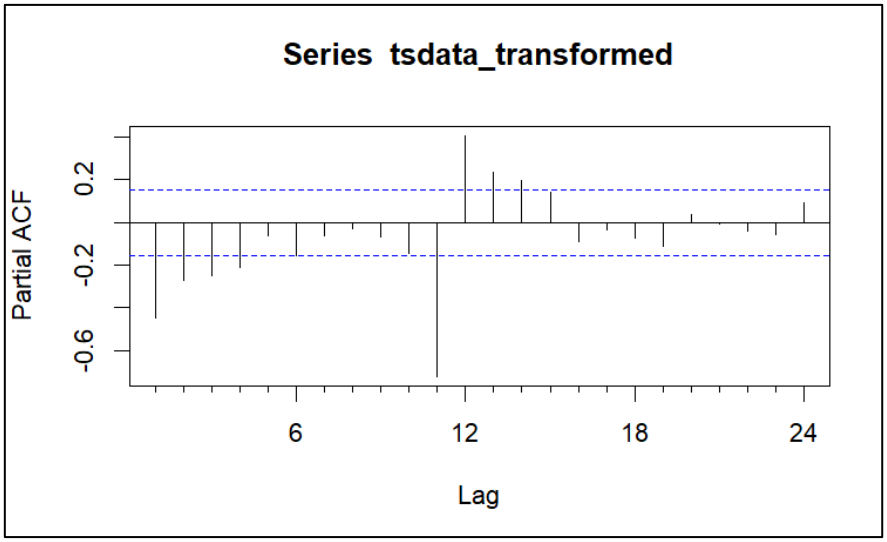

Fig 4: PACF plot of transformed timeseries

i. The ACF plot is used to determine the Moving Averages (MA) component value of the order of SARIMA model.

ii. The PACF plot is used to determine the Auto Regressive (AR) component of the SARIMA model.

iii. Seasonal MA(Q): As seen in the ACF plot fig above there are significant spikes at lag 12, 24, 36, 48, and other multiples of 12 which indicates seasonality. Because of the repeated spikes at multiples of 12 it suggests Q = 1.

iv. Non seasonal MA(q): There is a significant spike at lag 1 and the ACF cuts off after that the non-seasonal component will come as 1 i.e. q = 1.

v. Non seasonal AR(p): As seen in the above PACF plot there are 4 significant spikes till lag 4 and after that all the lags fall in the confidence interval which tells that the value of non-seasonal p should be 4.

vi. Seasonal AR(P): The PACF plot does show a spike at 12 which is just significant. Therefore, either value of P = 1 or P = 0 will be tried.

The data has been differenced once to remove the regular trend component and one seasonal differencing to remove the seasonal component. So, d = 1 and D = 1.

Considering the above investigation and analysis, the order of SARIMA model for monthly data will be

(4, 1, 1) (1, 1, 1) [12]

Suggest your candidate model together with a few possible variations. Describe how and on what bases you have selected your final model.

STAGE 2. Estimation

Validation of the selected initial candidate model

i. Candidate Model 1: ARIMA (4, 1, 1) (1, 1, 1) [12]

```
> fit1 <- Arima(log_tsdata, order=c(4,1,1), seasonal = c(1,1,1), include.constant = FALSE)
> summary(fit1)
Series: log_tsdata
ARIMA(4,1,1)(1,1,1)[12]

Coefficients:
          ar1      ar2      ar3      ar4      ma1    sar1     sma1
      -0.6383  -0.4466  -0.0604  -0.0316  -0.1186  0.1907  -0.9998
s.e.   0.9601   0.7276   0.5045   0.1215   0.9576  0.0884   0.1461

sigma^2 = 0.002098:  log likelihood = 241.03
AIC=-466.06   AICc=-465.04   BIC=-441.92

Training set error measures:
                      ME       RMSE        MAE         MPE     MAPE      MASE         ACF1
Training set -0.003070591 0.04291999 0.03152078 -0.1121407 1.158706 0.3656159 -0.006526053
> checkresiduals(fit1)

        Ljung-Box test

data:  Residuals from ARIMA(4,1,1)(1,1,1)[12]
Q* = 37.389, df = 17, p-value = 0.00298

Model df: 7.   Total lags used: 24
```

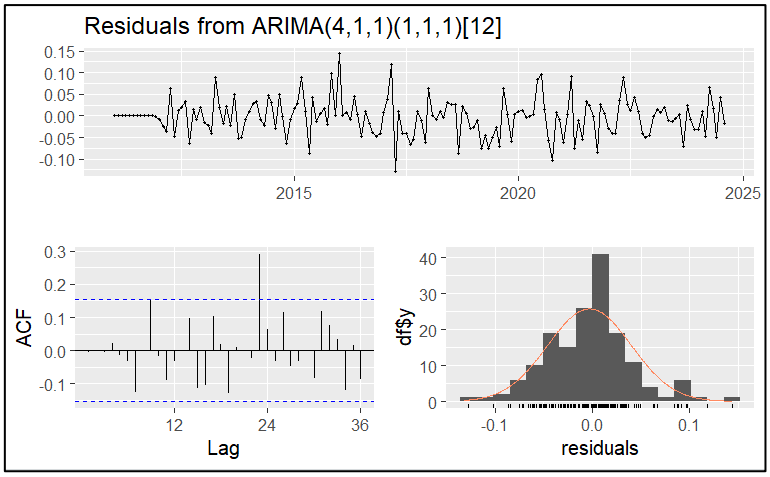

Fig 5: Residual plot for candidate model 1

From the Fig 5, the AIC value of -466.06 is reasonably low but not the best. Higher BIC value indicates overfitting. The standard errors are high which suggests some uncertainty in the coefficients. The residuals fluctuate randomly around the zero line however there are few extreme spikes. The histogram shows that the residuals are normally distributed with a peak affecting it.

Possible Variations for candidate model

ii. Candidate Model 2: ARIMA (4, 1, 1) (0, 1, 1) [12]

```
> fit2 <- Arima(log_tsdata, order=c(4,1,1), seasonal = c(0,1,1), include.constant = FALSE)
> summary(fit2)
Series: log_tsdata
ARIMA(4,1,1)(0,1,1)[12]

Coefficients:
          ar1      ar2      ar3      ar4      ma1     sma1
      -0.6700  -0.4587  -0.0628  -0.0403  -0.1102  -0.8213
s.e.   0.8358   0.6545   0.4469   0.1099   0.8327   0.1156

sigma^2 = 0.002339:  log likelihood = 239.1
AIC=-464.2   AICc=-463.42   BIC=-443.08

Training set error measures:
                      ME       RMSE        MAE         MPE    MAPE      MASE          ACF1
Training set -0.002976968 0.04547313 0.03331451 -0.1090011 1.22406 0.3864218 -0.005231102


> checkresiduals(fit2)

        Ljung-Box test

data:  Residuals from ARIMA(4,1,1)(0,1,1)[12]
Q* = 38.608, df = 18, p-value = 0.003215

Model df: 6.   Total lags used: 24
```

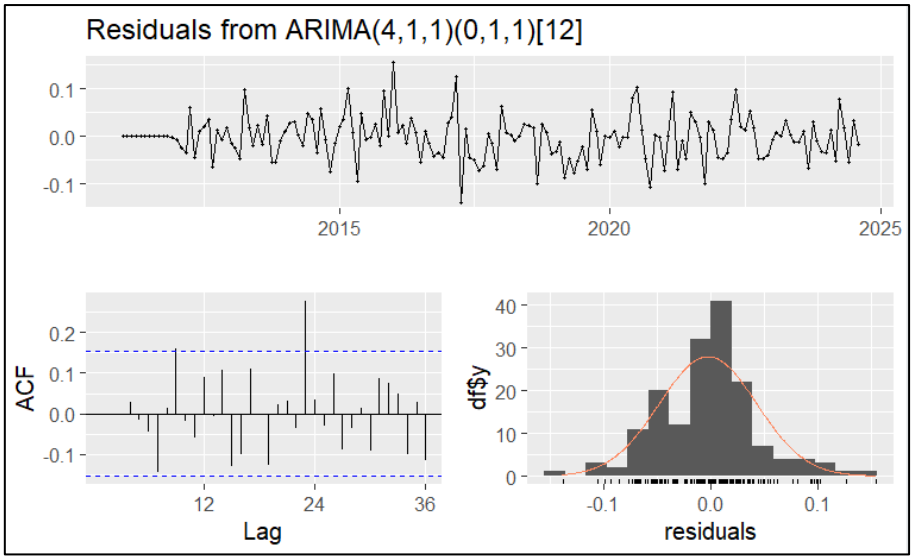

Fig 6: Residual plot for candidate model 2

In Fig 6, the residuals are fluctuating around zero showing no pattern which shows white noise. The AIC value of -464.2 is high as compared to the first candidate model. The histogram of residuals shows a bell curve. Also, majority of the lags in the ACF plot are inside the confidence level. The model is overall a good fit and better than the candidate model 1.

iii. Candidate Model 3 - ARIMA (4, 1, 0) (0, 1, 1) [12]

```
> fit3 <- Arima(log_tsdata, order=c(4,1,0), seasonal = c(0,1,1), include.constant = FALSE)
> summary(fit3)
Series: log_tsdata
ARIMA(4,1,0)(0,1,1)[12]

Coefficients:
          ar1      ar2      ar3      ar4     sma1
      -0.7798  -0.5434  -0.1192  -0.0484  -0.8213
s.e.   0.0828   0.1047   0.1043   0.0820   0.1156

sigma^2 = 0.002323:  log likelihood = 239.09
AIC=-466.18   AICc=-465.6   BIC=-448.08

Training set error measures:
                      ME       RMSE        MAE         MPE     MAPE      MASE         ACF1
Training set -0.002956463 0.04547551 0.03329189 -0.1082984 1.22335 0.3861594 -0.00605123

> checkresiduals(fit3)

        Ljung-Box test

data:  Residuals from ARIMA(4,1,0)(0,1,1)[12]
Q* = 38.745, df = 19, p-value = 0.004765

Model df: 5.   Total lags used: 24
```

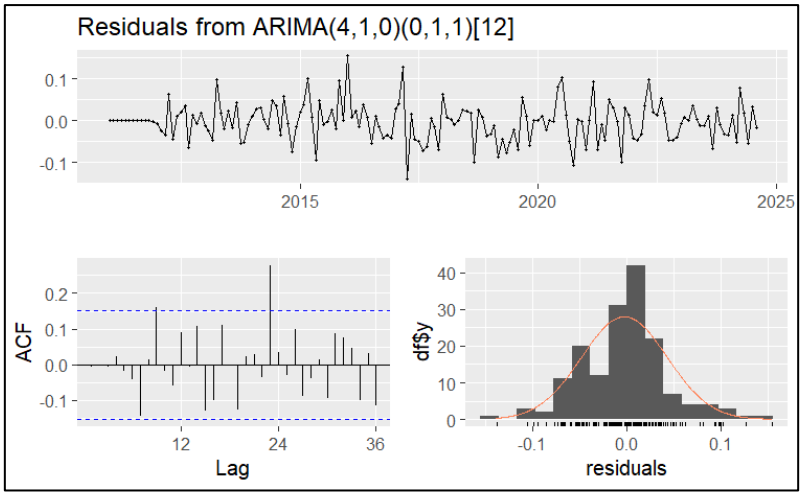

Fig 7: Residual plot for candidate model 3

A lower AIC value of -466.18 and BIC of 448.08 indicates a much better fit than above 2 models. The standard errors are low which suggests the terms are significant.

The bell-shaped curve of histogram as seen in Fig 7 shows that the residuals are normally distributed. Also, the residuals are fluctuating around zero without a seasonal or trend pattern which suggests white noise.

STAGE 3. Diagnostics

Statistic/Model	fit1 (ARIMA(4,1,1)(1,1,1)[12])	fit2 (ARIMA(4,1,1)(0,1,1)[12])	fit3 (ARIMA(4,1,0)(0,1,1)[12])
Log Likelihood	241.03	239.1	239.09
AIC	-466.06	-464.2	-466.18
AICc	-465.04	-463.42	-465.6
BIC	-441.92	-443.08	-448.08
ME	-0.003070591	-0.002976968	-0.002956463
RMSE	0.04291999	0.04547313	0.04547551

Considering the above analysis and lower AIC and BIC and the Model 3 ARIMA (4,1,0) (0,1,1) [12] appears to be a best fit from the above variations.

Also, Candidate model 3 has the highest p-value from the Ljung-Box test from all of the models which supports the analysis.

Forecasting with ARIMA (4,1,0) (0,1,1) [12] model.

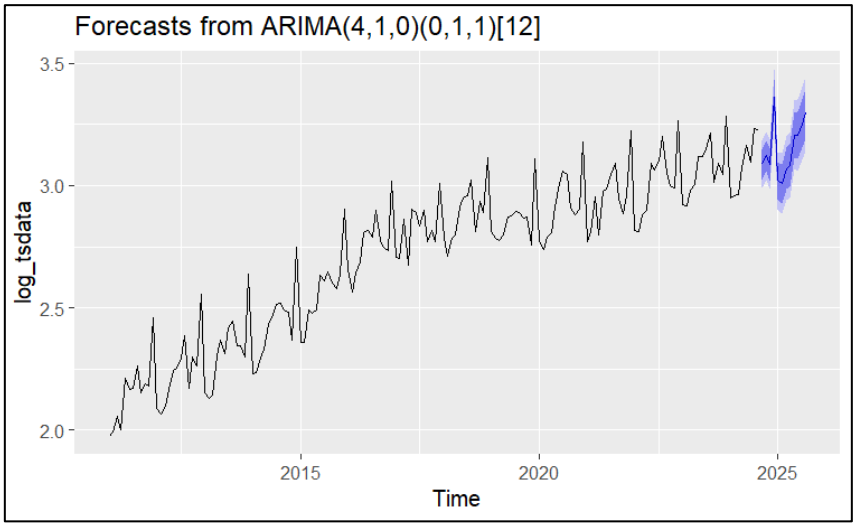

Fig 9: Forecasted Movin timeseries with best fir ARIMA model

The forecast appears to follow the trend and seasonality pattern of the old(historical) data over the years. The confidence intervals (the blue area around the black forecast line) is little which suggest that the forecasting is doing good with this ARIMA model. Also, tight confidence intervals tells that there is not much uncertainty in the forecasting.

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Sep 2024	3.086108	3.024280	3.147936	2.991551	3.180666
Oct 2024	3.121232	3.057924	3.184541	3.024411	3.218054
Nov 2024	3.085610	3.019897	3.151323	2.985111	3.186110
Dec 2024	3.362529	3.288851	3.436207	3.249848	3.475210
Jan 2025	3.020208	2.943420	3.096997	2.902770	3.137647
Feb 2025	3.007364	2.926892	3.087836	2.884293	3.130436
Mar 2025	3.069869	2.985161	3.154577	2.940319	3.199419
Apr 2025	3.081461	2.993443	3.169478	2.946850	3.216072
May 2025	3.205930	3.114466	3.297394	3.066048	3.345813
Jun 2025	3.205202	3.110356	3.300048	3.060147	3.350256
Jul 2025	3.252908	3.154927	3.350889	3.103058	3.402757
Aug 2025	3.294565	3.193476	3.395654	3.139963	3.449167

Table 1: Forecasted Values

Comparing the model with previous forecasting done in Section 1, the confidence interval levels are narrower of the ARIMA model. The model has multiple AR components and seasonal MA which tend to capture the trends and seasonality in a better and flexible way.

Overall low variability and better accuracy is achieved by using the ARIMA model in the forecasts than the ETS forecast.

APPENDIX

```
library(readxl)
library(forecast)
library(seasonal)
library(seasonalview)
library(ggplot2)
library(lmtest)


tsdata = ts(trend, start = c(2011,1), frequency = 12)

autoplot(tsdata)
m_acf <-Acf(tsdata,lag=NULL)
autoplot(m_acf)

tsdisplay(diff(tsdata))

m_pacf <-Pacf(tsdata,lag=50)
autoplot(m_pacf)

#----------stationarity----------#

tsdata = ts(trend, start = c(2011,1), frequency = 12)
plot(tsdata)

m_acf <-Acf(tsdata,lag=100)
autoplot(m_acf) #tells the series is non stationary

#pacf after diff
pacf <- Pacf(tsdata,lag=50) #P seems to be 1, D = 1

#transforming
log_tsdata <- log(tsdata)
autoplot(log_tsdata)

#Differencing the transformed data to remove trend
d <- diff(log_tsdata, lag = 1)
#d <- diff(d, lag =12)
autoplot(d)

#boxcox to remove seasonality
bc_trans <- BoxCox.lambda(d)
tsdata_transformed <- BoxCox(d, bc_trans)
autoplot(tsdata_transformed)

# Fit an AR model to the transformed data
fit <- ar(tsdata_transformed)
residuals_squared <- residuals(fit)^2

# Perform the Ljung-Box test on the squared residuals
```

```
Box.test(residuals_squared, lag = 12, type = "Ljung-Box")


#acf after diff
Acf(tsdata_transformed,lag=100)

#pacf after diff
pacf_tfd <- Pacf(tsdata_transformed,lag=24)
plot(pacf_tfd, main = "Movin transformed PACF")

#order of arima 1st candidate model
fit1 <- Arima(log_tsdata, order=c(4,1,1), seasonal = c(1,1,1), include.constant = FALSE)
summary(fit1)
checkresiduals(fit1)

#order of arima 2nd candidate model
fit2 <- Arima(log_tsdata, order=c(4,1,1), seasonal = c(0,1,1), include.constant = FALSE)
summary(fit2)
checkresiduals(fit2)

#order of arima 3rd candidate model
fit3 <- Arima(log_tsdata, order=c(4,1,0), seasonal = c(0,1,1), include.constant = FALSE)
summary(fit3)
checkresiduals(fit3)

forecasts_sarima <- forecast(fit3, h=12)  # for example, forecast 12 periods ahead
autoplot(forecasts_sarima)
summary(forecasts_sarima)
forecast_df <- as.data.frame(forecasts_sarima)
```