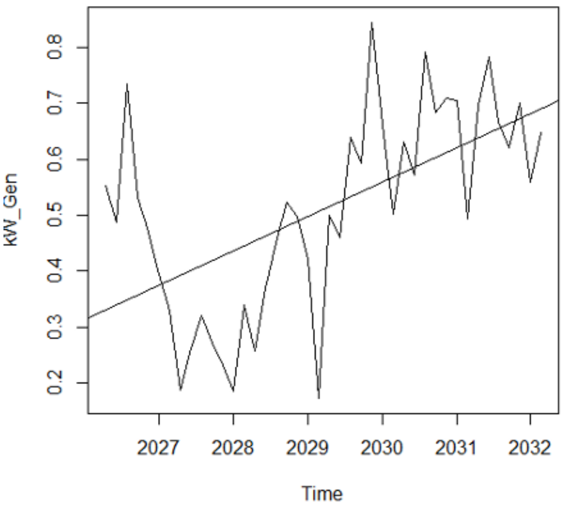
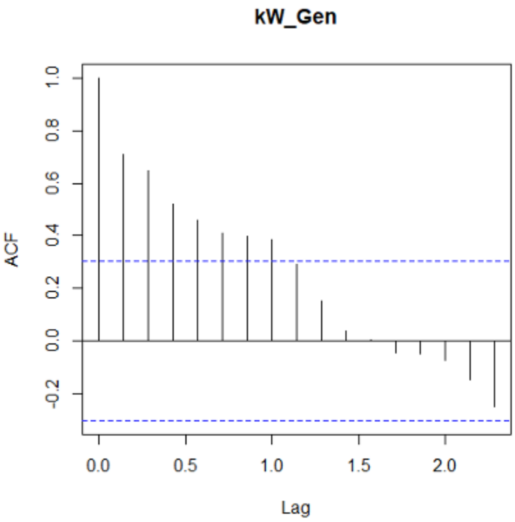
```
#Following packages are essential for Time Series forecasting
install.packages("fUnitRoots")
install.packages("lmtest")
install.packages("FitAR")
install.packages("forecast")
install.packages("haven")
install.packages("lmtest")
install.packages("tseries")

#Read the packages
library(haven)
library(fUnitRoots)
library(lmtest)
library(FitAR)
library(forecast)
library(lmtest)
library(tseries)

#Set the working directory
dir = 'C://OSU 2019-2021//Semester - 3//BAN 5753//Week14'
setwd(dir)
getwd()

#Read SAS data
df = read_sas('solarpv.sas7bdat')

#Understanding data
head(df,15)              #Top 15 records
summary(df)              #Summary of the data
nrow(df)                 #Number of rows
ncol(df)                 #Number of columns
names(df)                #Names of the columns
class(df$EDT)            #Data Type of EDT
class(df$kW_Gen)         #Data Type of Power Generation
```

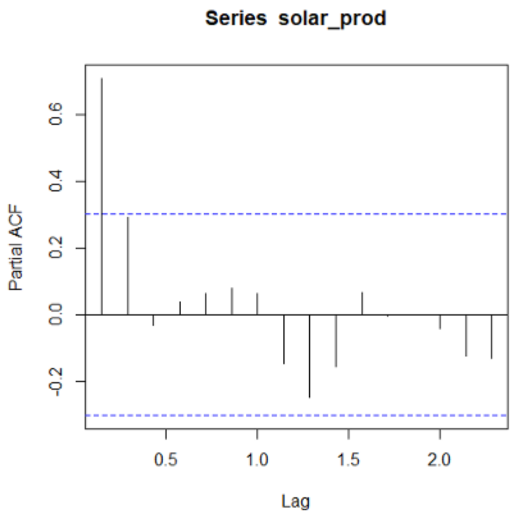
```
> head(df,15)                     #Top 15 records
# A tibble: 15 x 4
     EDT kW_Gen Cloud_Cover cosval
   <dbl>  <dbl>       <dbl>  <dbl>
 1 20001  0.553        4.75 -0.301
 2 20008  0.487        5.34 -0.413
 3 20015  0.734        2.29 -0.519
 4 20022  0.531        4.92 -0.618
 5 20029  0.471        5.52 -0.708
 6 20036  0.394        5.72 -0.788
 7 20043  0.330        5.02 -0.856
 8 20050  0.188        6.57 -0.912
 9 20057  0.262        6.03 -0.954
10 20064  0.320        4.52 -0.983
11 20071  0.273        5.20 -0.998
12 20078  0.232        6.27 -0.998
13 20085  0.185        6.40 -0.983
14 20092  0.339        4.49 -0.954
15 20099  0.258        6.08 -0.912
> summary(df)                     #Summary of the data
      EDT            kW_Gen         Cloud_Cover         cosval
 Min.   :20001   Min.   :0.1730   Min.   :2.286   Min.   :-0.99759
 1st Qu.:20073   1st Qu.:0.3753   1st Qu.:4.591   1st Qu.:-0.78784
 Median :20145   Median :0.5124   Median :5.288   Median :-0.30064
 Mean   :20145   Mean   :0.5111   Mean   :5.190   Mean   :-0.08111
 3rd Qu.:20216   3rd Qu.:0.6592   3rd Qu.:5.821   3rd Qu.: 0.76241
 Max.   :20288   Max.   :0.8446   Max.   :6.571   Max.   : 0.99808
> nrow(df)                        #Number of rows
[1] 42
> ncol(df)                        #Number of columns
[1] 4
> names(df)                       #Names of the columns
[1] "EDT"         "kW_Gen"        "Cloud_Cover" "cosval"
> class(df$EDT)                   #Data Type of EDT
[1] "numeric"
> class(df$kW_Gen)
[1] "numeric"
```

```
#Subsetting the data to include
solar_prod = subset(df,select = c("kW_Gen"))

#Converting EDT to time format
df$EDT = as.Date(df$EDT, origin = "1970-01-01")
print(head(df$EDT,5))       #First recorded date for power generation
print(tail(df$EDT,5))       #Last recorded date for power generation

#Convert solar production to timeseries data with origin as 2025-10-05 with 7 day interval
solar_prod = ts(solar_prod,frequency = 7, start = c(2025,10,05))

#Plot the trend with a mean line
plot(solar_prod)
abline(reg=lm(solar_prod~time(solar_prod)))
```




```
#Calculate ACF for the timeseries
acf(solar_prod)

#Calculate PACF for the timeseries
pacf(solar_prod)
```





#LjungBox Chi Square test to check for White noice in the timeseries
Box.test(solar_prod, type="Ljung-Box")

```
        Box-Ljung test

data:  solar_prod
X-squared = 22.669, df = 1, p-value = 1.925e-06
```

#Create a ARIMAx with (p=1, d=0, q=0) AND 2 independent variables
#We are using Arima from the forecast package and not arima from the stat package
#ARIMAX forecast/prediction won't work on the stat package
#Create a dataframe named ind_data which will include all independent variables

ind_data=cbind('cloud_cover'=df1$Cloud_Cover,'cosval'=df1$cosval)

#The difference between Arima and Arimax is the presence of independent X variable, the variables are added in the model under xreg
fitARIMAx <- Arima(df1$kW_Gen, order=c(1,0,0),method="ML",xreg=ind_data)

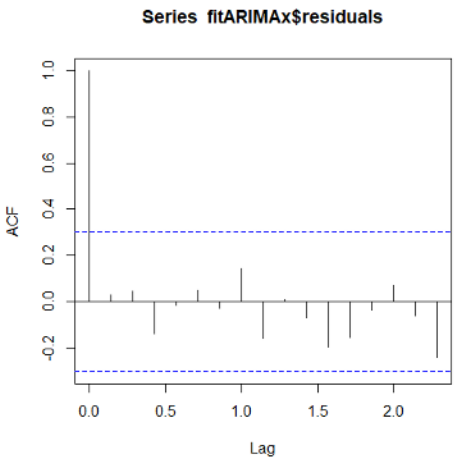
fitARIMAx #Print all detials of the ARIMA Model
summary(fitARIMAx)

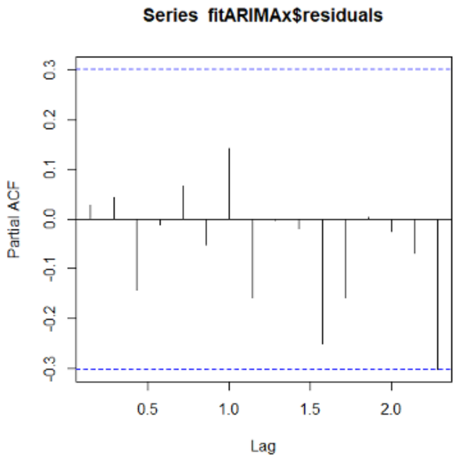
```
Series: df1$kW_Gen
Regression with ARIMA(1,0,0) errors

Coefficients:
         ar1  intercept  cloud_cover  cosval
      0.5557     0.9974      -0.0909  0.1666
s.e.  0.1249     0.0548       0.0097  0.0285

sigma^2 estimated as 0.0045:  log likelihood=55.8
AIC=-101.6   AICc=-99.93   BIC=-92.91

Training set error measures:
                        ME       RMSE        MAE       MPE     MAPE      MASE       ACF1
Training set -0.0006524516 0.06380786 0.04973271 -2.830267 11.65273 0.2950455 0.02904214
```

coeftest(fitARIMAx) #Check coefficient of the ARIMA Model

```
z test of coefficients:

             Estimate Std. Error z value  Pr(>|z|)    
ar1          0.5557209  0.1249097  4.4490 8.628e-06 ***
intercept    0.9974417  0.0548140 18.1969 < 2.2e-16 ***
cloud_cover -0.0909047  0.0097468 -9.3266 < 2.2e-16 ***
cosval       0.1666422  0.0285226  5.8425 5.143e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

#Perform aCF, PACF and Ljung Box (White Noise) test - We will use the residuals for this
acf(fitARIMAx$residuals)
pacf(fitARIMAx$residuals)

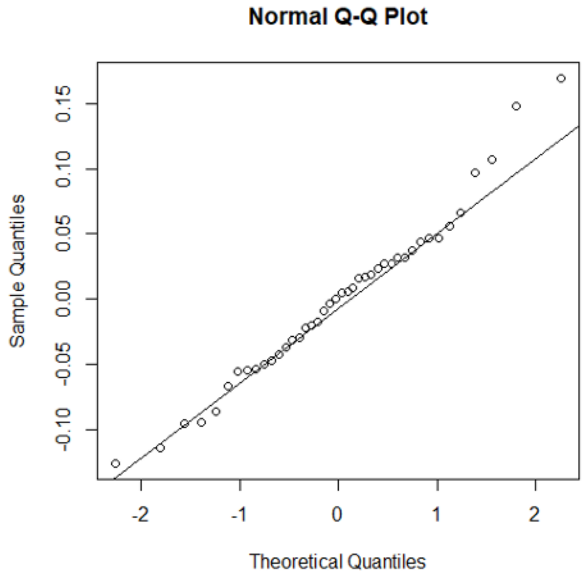




Box.test(fitARIMAx$residuals, type="Ljung-Box")

```
        Box-Ljung test

data:  fitARIMAx$residuals
X-squared = 0.038017, df = 1, p-value = 0.8454
```

qqnorm(fitARIMAx$residuals)
qqline(fitARIMAx$residuals)



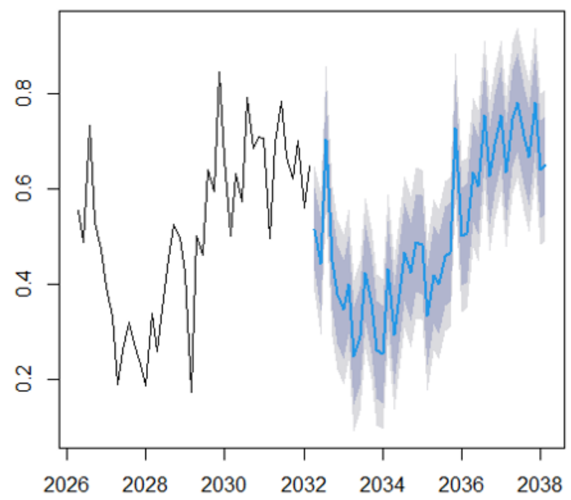
#Diagnostic Checking
arimax_bic = AIC(fitARIMAx , k = log(length(df1$kW_Gen)))
print(arimax_bic)

#Forecast for the next 5 weeks based on 42 weeks
arimax_fore = forecast(fitARIMAx,xreg=ind_data, h = 5)
accuracy(arimax_fore)

```
                        ME       RMSE        MAE       MPE     MAPE      MASE       ACF1
Training set -0.0006524516 0.06380786 0.04973271 -2.830267 11.65273 0.2950455 0.02904214
```

plot(arimax_fore)

Forecasts from Regression with ARIMA(1,0,0) errors

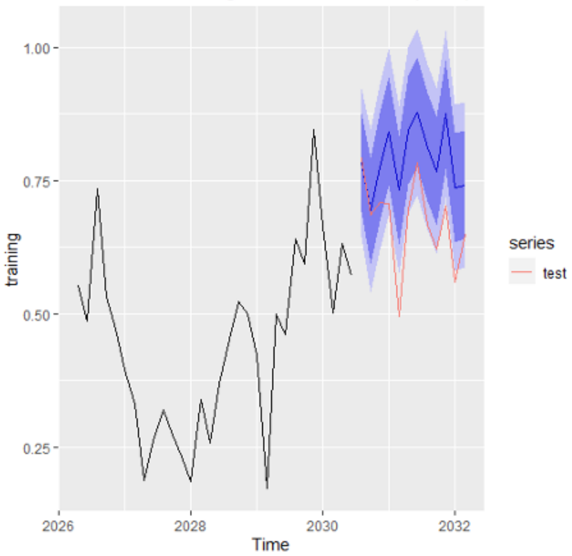


#Training and Testing for time series using multiple time series
training <- subset(df1$kW_Gen, end=length(df1$kW_Gen)-12) #Creating a training data set 1-30
test <- subset(df1$kW_Gen, start=length(df1$kW_Gen)-11) #Creating a test data set 31-42
#Splitting the data for the independent variable in the same way as the time series data
ind_train <-head(ind_data,30)
ind_test <-tail(ind_data,12)
#Confirming the dimension of time series and independent data, it should be same
length(training)
nrow(ind_train)

solar_prod_ts <- Arima(training, order=c(1,0,0),method="ML",xreg=ind_train) #Arima Model with p=1

#Plot train + Test on the graph
solar_prod_ts %>%
 forecast(h=12,xreg=ind_test) %>%
 autoplot() + autolayer(test)

Forecasts from Regression with ARIMA(1,0,0) errors



#Check the accuracy of the model on the test data
solar_prod_ts_test <- Arima(test, model=solar_prod_ts,xreg=ind_test)
accuracy(solar_prod_ts_test)

```
                     ME       RMSE        MAE       MPE     MAPE      MASE      ACF1
Training set -0.07297406 0.08682121 0.07297406 -11.70097 11.70097 0.7372044 0.1064228
```

#The above output is for the test dataset, R Script labels it as a training set