Math1318_Time Series_Assignment_2

2022-04-24

Task:

Executive Summary:

This report is about analyzing the yearly changes in Antarctica land ice mass in billion metric tons relative to the ice mass in 2001. We analysed the data, given descriptive analysis, checking for stationary and non stationary, if data is non- stationary, applied transformation on the data wherever required, plotted the ACF & PACF along with EACF & BIC table. Moreover, proposed a set of possible ARIMA(p, d, q) models using all suitable model specification tools.

Table of Contents

Introduction:

Information about the dataset:

The dataset consist of yearly changes in Antarctica land ice mass in billion metric tons relative to the ice mass in 2001 (Source: https://www.epa.gov/climate-indicators/climate-change-indicators-ice-sheets (https://www.epa.gov/climate-indicators/climate-change-indicators-ice-sheets) (Links to an external site.)). A negative value in the dataset represents a decrease in the ice mass from the level in 2001 and a positive value represents an increase in the ice mass over the level in 2001.

Loading Data:

```
# Required packages for this task:
library(readr)  # for reading csv file
library(TSA)    # for time series data
```

```
##
## Attaching package: 'TSA'
```

```
## The following object is masked from 'package:readr':
##
##     spec
```

```
## The following objects are masked from 'package:stats':
##
##     acf, arima
```

```
## The following object is masked from 'package:utils':
##
##     tar
```

```
library(tseries) # for for Automatic Time Series Forecasting:
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
## Load data :
df <- read.csv("assignment2Data2022.csv")
# converting the data in time series, its a continues annualy data statring from 2002
and ending on 2020, so using frequency =1.
#Time Series:
data<- ts(df$NASA...Annual.Antarctica.land.ice.mass, start = 2002, end=2020 , frequen
cy = 1)
# checking the class of the data
class(data)
```

```
## [1] "ts"
```

```
head(data) # for showing top 6 obsevations
```

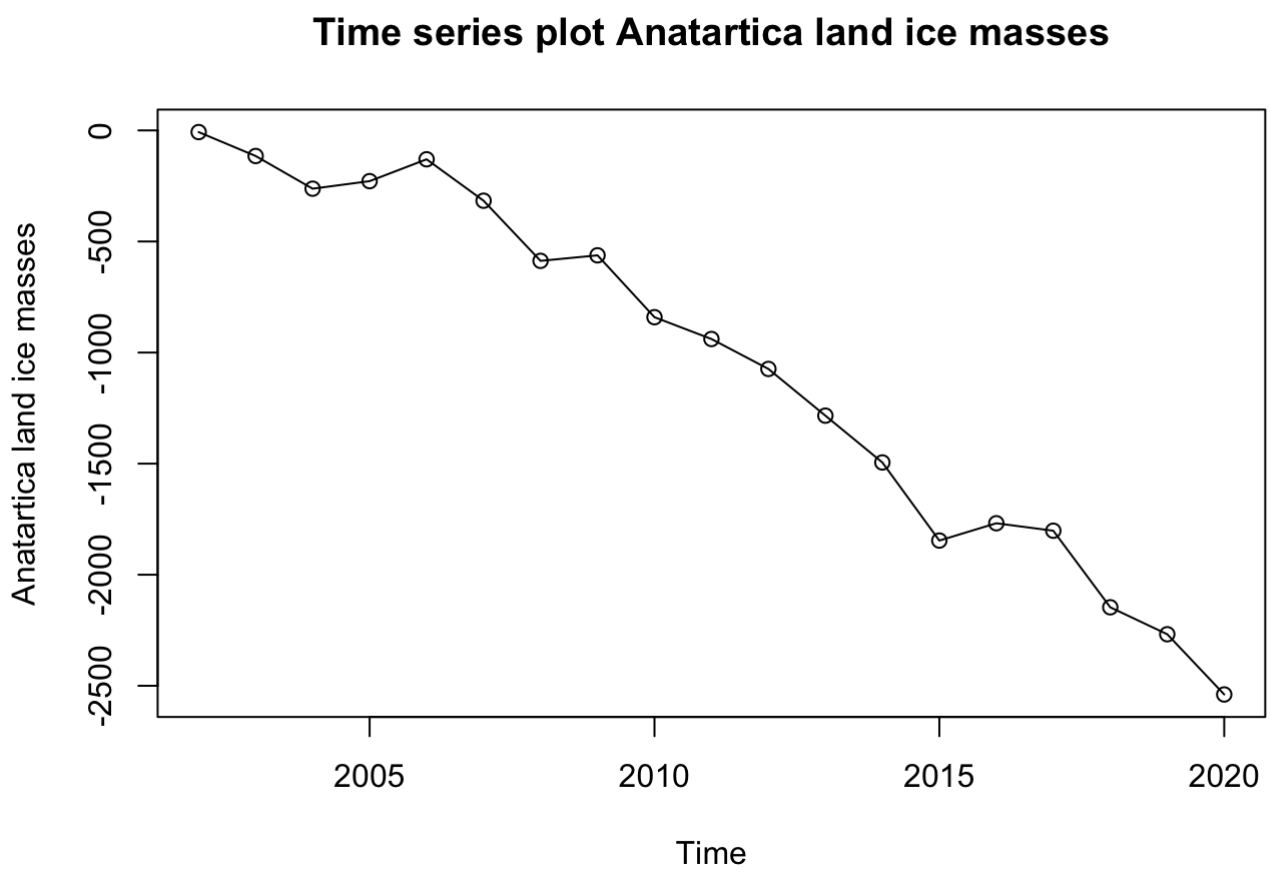
```
## [1]   -7.546667 -115.186364 -262.302500 -228.376667 -129.987500 -316.476667
```

Descriptive Analysis:

3.2 Analyzing the Data:

Data is in time series, Let's plot the graph and observe the trend, seasonality, changing variance and behavior of the time series.

```
par(mfrow=c(1,1))
plot(data,type='o',ylab="Anatartica land ice masses", main='Time series plot Anatarti
ca land ice masses')
```




```
summary(data)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -2539.055 -1785.137  -938.991 -1063.774  -289.390    -7.547
```

After converting the data to time series format , we generated a time series plot and analyzed the following :

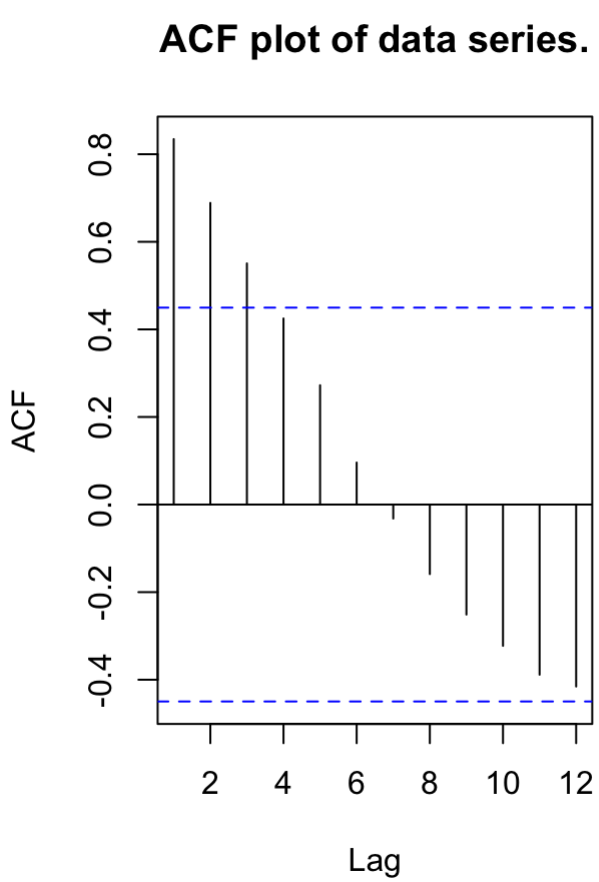
- Trend :We can see a clear downward trend.
- Seasonality: There is No seasonality, a seasonal pattern is rise and fall in the data values that repeats after regular time intervals.
- No obvious intervention.
- Changing Variance : no sudden increasing and decreasing variance, so no Change in Variance.
- Behavior : Series appeared to be auto regressive due to multiple succeeding observations.

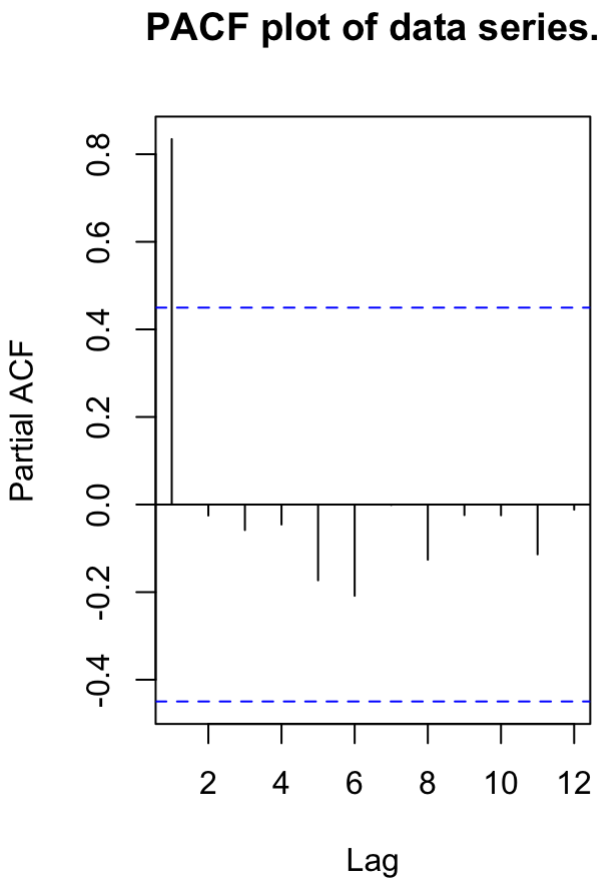
3.2.1 ACF & PACF Plots:

- **ACF** Plots are designed to show whether the elements of time series are positively correlated, negatively correlated or independent on each other.We calculate the correlation for time series observations with previous time stamps called lags
- **PACF** A partial auto correlation is the amount of correlation between a variable and a lag of itself that is not explained by correlations at all lower-order-lags.

Output of ACF and PACF: Slowly decaying pattern in ACF and very high first correlation in PACF implies the existence of trend and nonstationarity.

Let's go for other tests:





ADF Test:

We then perform Augmented Dickey Fuller (ADF) test to check whether the given time series is stationary or not.It is one of the most commonly used test when it comes to analyzed the stationarity of data.(if p-value is > significance level (say 0.05), then the series is non-stationary, here we get p = 0.3004, so implies that series is non-stationary)

```
adf.test(data)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  data
## Dickey-Fuller = -2.7141, Lag order = 2, p-value = 0.3004
## alternative hypothesis: stationary
```

```
# p value is greater than 0.05, need to look for other test
```

- **ADF test Output**: The p-value is obtained is greater than significance level of 0.05 and the ADF statistic is higher than any of the critical values. Clearly, there is no reason to reject the null hypothesis. So, the time series is in fact non-stationary.

PP Test:

The Phillips–Perron test is a unit root test. That is, it is used in time series analysis to test the null hypothesis that a time series is integrated of order 1.(if p-value is > significance level (say 0.05), then the series is non-stationary, here we get p = 0.7516, so implies that series is non-stationary)

```
pp.test(data)   # pp test
```

```
##
##  Phillips-Perron Unit Root Test
##
## data:  data
## Dickey-Fuller Z(alpha) = -5.8754, Truncation lag parameter = 2, p-value
## = 0.7516
## alternative hypothesis: stationary
```

- In **pp test**, p value is greater than 0.05, fail to reject the null hypothesis.

KPSS Test:

The Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test figures out if a time series is stationary around a mean or linear trend, or is non-stationary due to a unit root, for alpha levels of 10%, 5% and 1%), then the null hypothesis is rejected; the series is non-stationary.

```
kpss.test(data) # KPSS test
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  data
## KPSS Level = 0.72818, Truncation lag parameter = 2, p-value = 0.01098
```

- In KPSS **test**, That is, if p-value is < significance level (say 0.05), then the series is non-stationary, here we get p=0.01, so implies that series is non-stationary.

Shapiro-Wilk Test:

A Shapiro-Wilk test is the test to check the normality of the data. The null hypothesis for Shapiro-Wilk test is that your data is normal, and if the p-value of the test if less than 0.05, then you reject the null hypothesis at 5% significance and conclude that your data is non-normal.

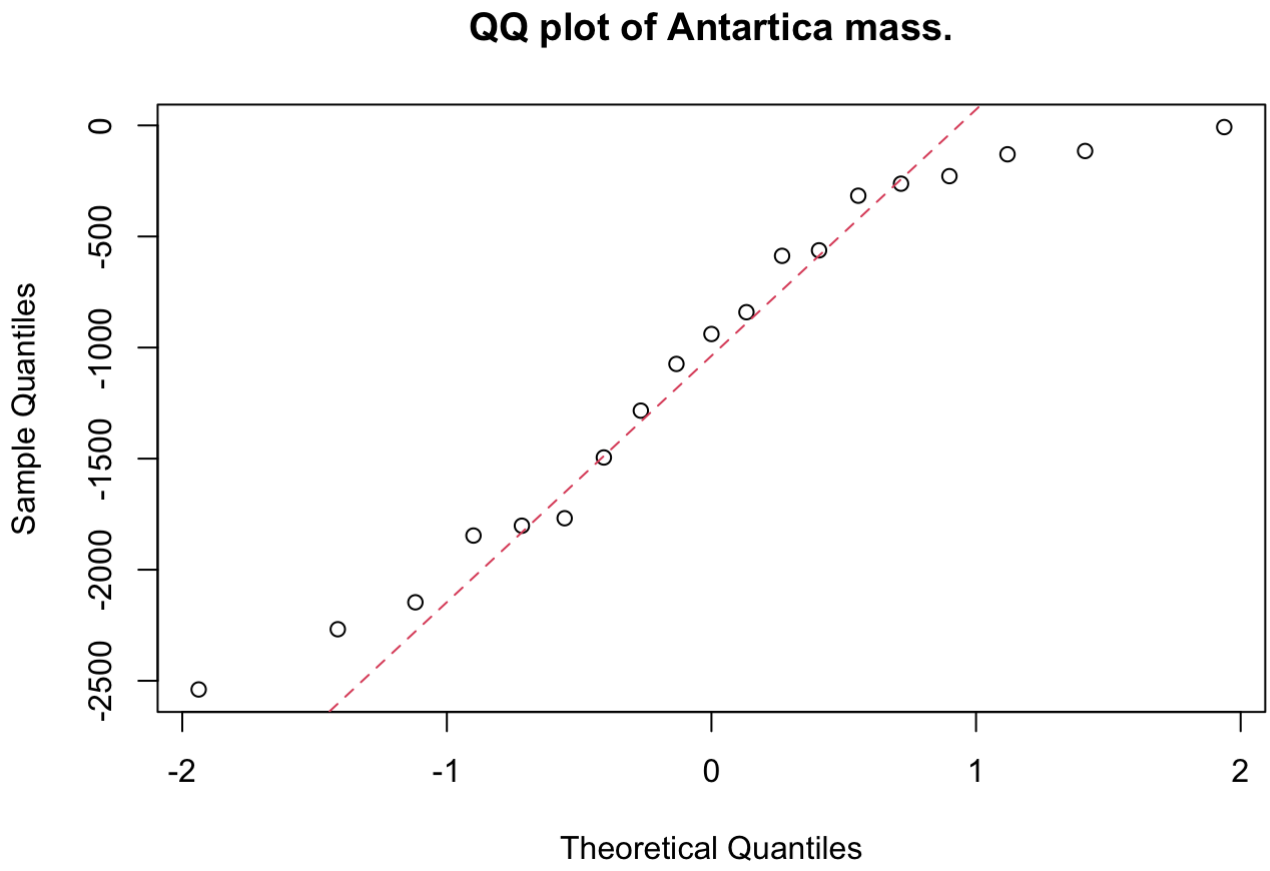
In **Shapiro-Wilk normality test**, p value is greater than 0.05, fail to reject the null hypothesis.

```
shapiro.test(data) # shapiro- Wilk test
```

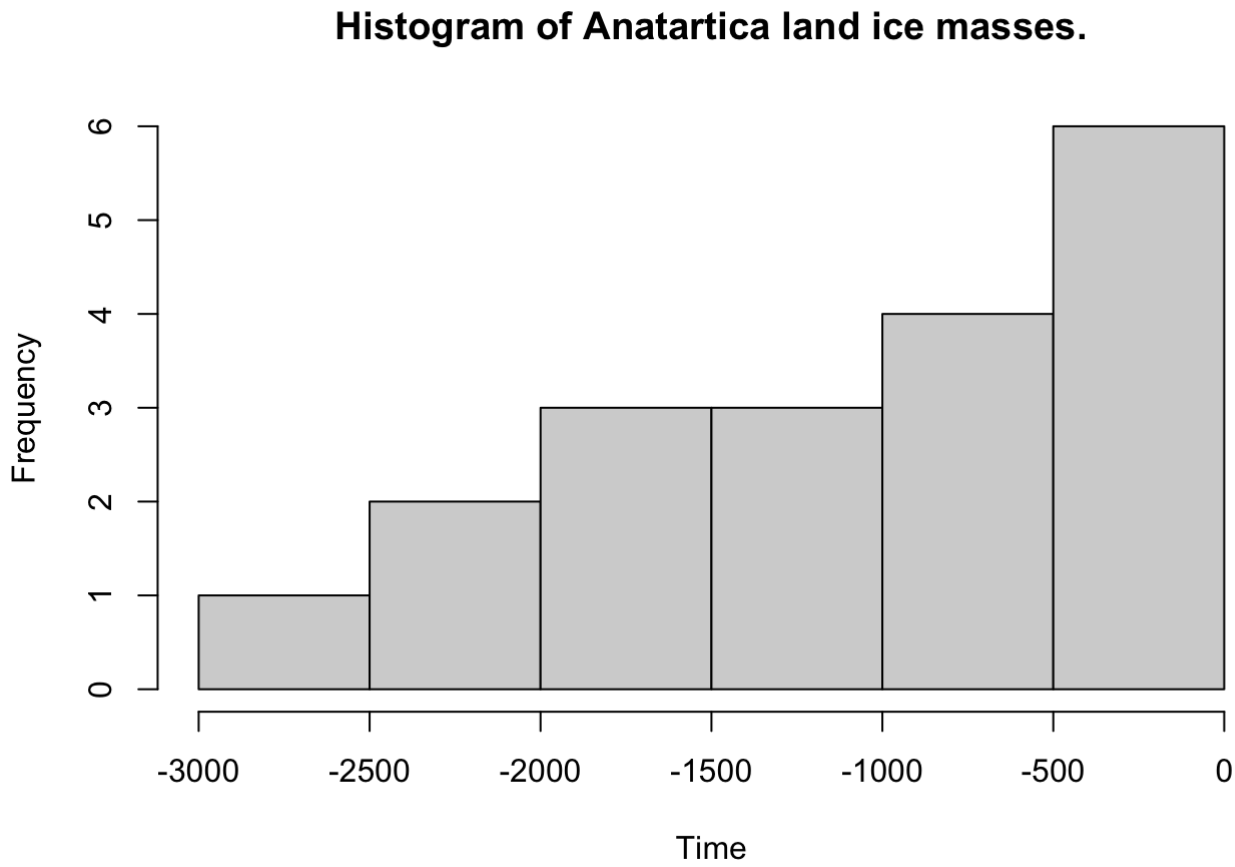
```
##
##  Shapiro-Wilk normality test
##
## data:  data
## W = 0.92616, p-value = 0.1471
```

QQ Plot and histogram:

```
# p value is greater than 0.05, need to look for other test
qqnorm(y=data, main = "QQ plot of Antartica mass.") # QQ plot
qqline(y=data, col = 2, lwd = 1, lty = 2)
```




```
hist(data,xlab='Time', main = "Histogram of Anatartica land ice masses.") # for check
ing normality
```



- **In QQ plot,** i can see some residuals, not along with normality, data is not normally distributed. Or, In the QQ plot the tails of the distribution is far from the normality. The histogram is left skewed, far from normality. OR the mean of a time series is not constant over time then it is non stationary. Hence, we have enough evidence to reject the normality hypothesis.
- As, series is non- stationary, we need do convert it into stationary, by using “differencing”.
- As data has negative observations, i can not apply box -cox transformation,
- We need to change the negative values into positive by adding the constant , so lets apply and improve normality of the observations.

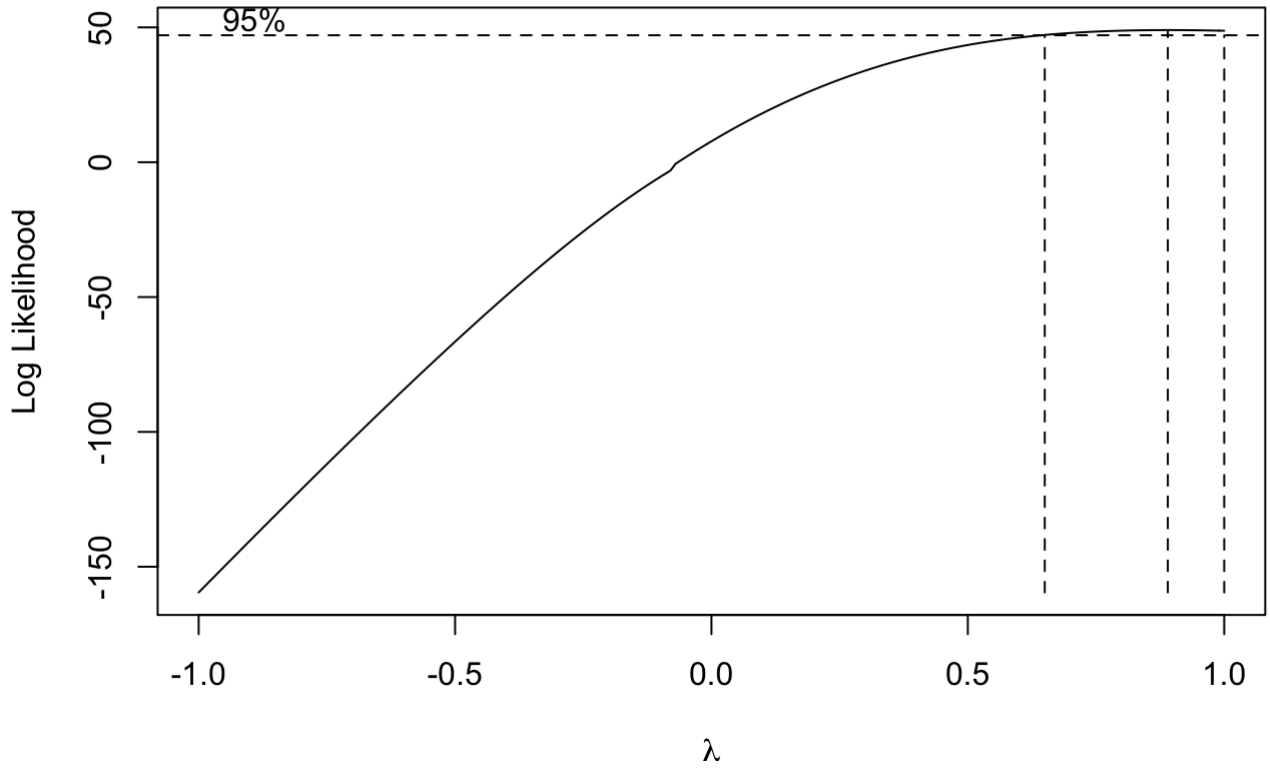
Boxcox Transformation :

A Box Cox transformation is a transformation of non-normal dependent variables into a normal shape. Normality is an important assumption for modeling, if the data isn’t normal, applying a Box-Cox means that you are able to run a broader number of tests.

```
#BC <- BoxCox.ar(data) # getting an error
#,lambda = seq(-1, 0.5, 0.01) If you get an error.
#adding constant and making data positive
data1<- data + abs(min(data)) + 0.01
#BC <- BoxCox.ar(data1)
BC <- BoxCox.ar(data1 ,lambda = seq(-1, 1, 0.01))
```

```
## Warning in arima0(x, order = c(i, 0L, 0L), include.mean = demean): possible
## convergence problem: optim gave code = 1

## Warning in arima0(x, order = c(i, 0L, 0L), include.mean = demean): possible
## convergence problem: optim gave code = 1
```




```
BC$ci
```

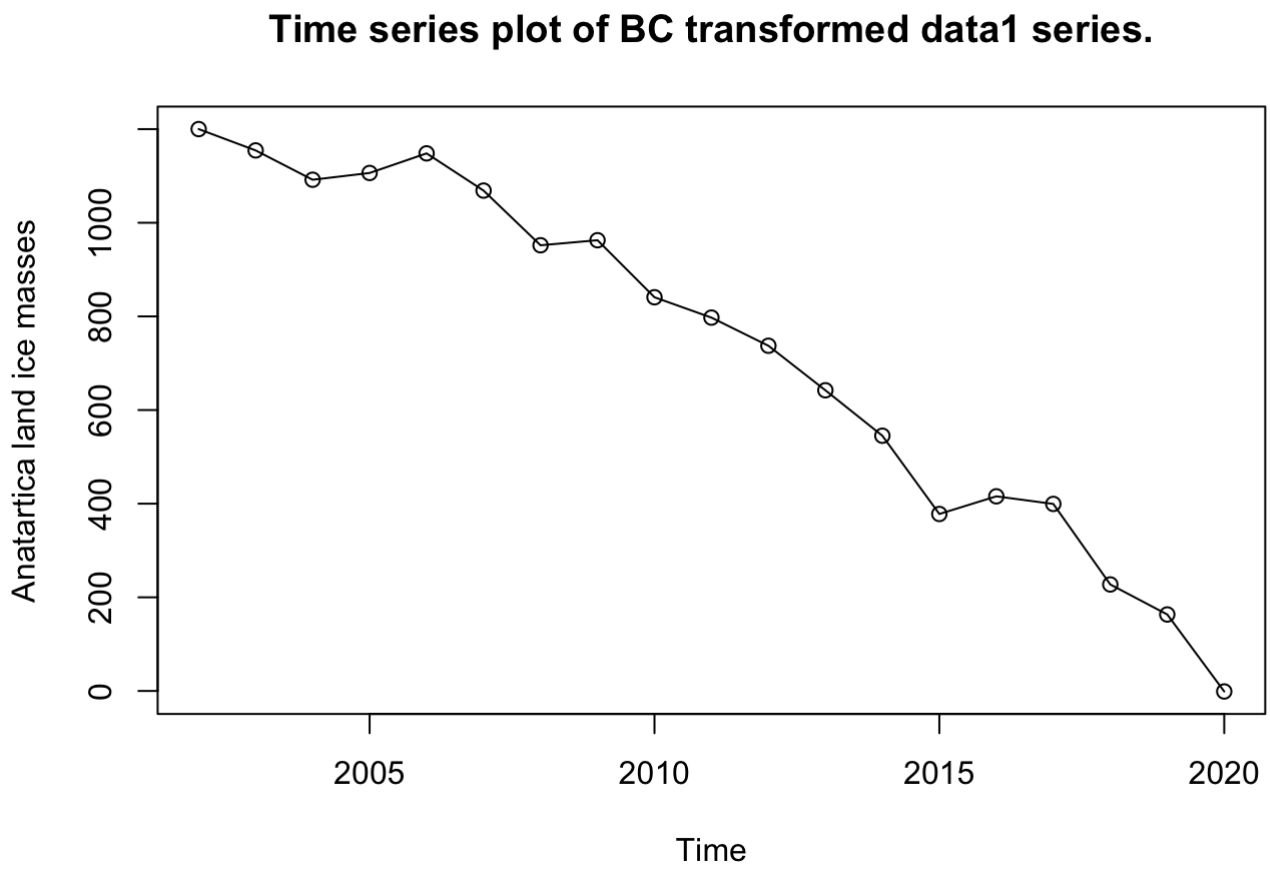
```
## [1] 0.65 1.00
```

```
lambda <- BC$lambda[which(max(BC$loglike) == BC$loglike)]
lambda
```

```
## [1] 0.89
```

```
BC.data1 = ((data1^lambda)-1)/lambda

par(mfrow=c(1,1))
plot(BC.data1,type='o', ylab ="Anatartica land ice masses", main="Time series plot of
BC transformed data1 series.")
```



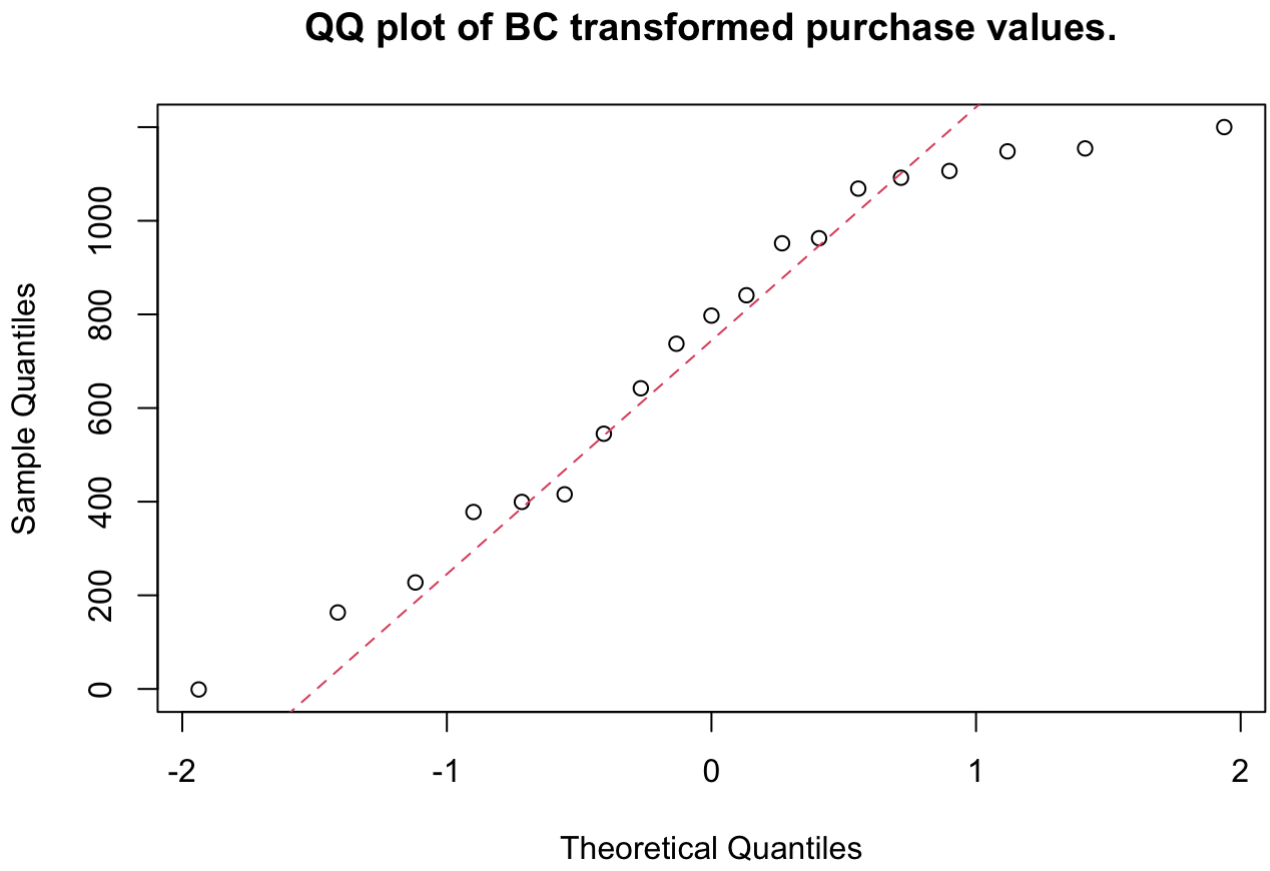
After applying box-cox, i can observe by QQ plot and ADF, pp and KPSS test,

- **QQ plot,** i can see some residuals, not along with normality, data is not normally distributed.
- shapiro test having , p value is greater than 0.05, fail to reject the null hypothesis.
- In Augmented Dickey-Fuller Test The p-value is obtained is greater than significance level of 0.05 and the ADF statistic is higher than any of the critical values. Clearly, there is no reason to reject the null hypothesis. So, the time series is in fact non-stationary.
- In Phillips-Perron Unit Root Test, The p-value is obtained is greater than significance level of 0.05 and the ADF statistic is higher than any of the critical values. Clearly, there is no reason to reject the null hypothesis. So, the time series is in fact non-stationary.

- Still, the series is not stationary:

- I need to look for differencing, so lets use differencing to improve normality of the observations.

```
# QQ Plot
qqnorm(y=BC.data1, main = "QQ plot of BC transformed purchase values.")
qqline(y=BC.data1, col = 2, lwd = 1, lty = 2)
```




```
# Kpss test:
kpss.test(data1)
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  data1
## KPSS Level = 0.72818, Truncation lag parameter = 2, p-value = 0.01098
```

```
# ADF test:
adf.test(data1)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  data1
## Dickey-Fuller = -2.7141, Lag order = 2, p-value = 0.3004
## alternative hypothesis: stationary
```

```
#pp test:
pp.test(data1)
```

```
##
##  Phillips-Perron Unit Root Test
##
## data:  data1
## Dickey-Fuller Z(alpha) = -5.8754, Truncation lag parameter = 2, p-value
## = 0.7516
## alternative hypothesis: stationary
```

```
# shapiro.test:
shapiro.test(BC.data1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  BC.data1
## W = 0.92582, p-value = 0.145
```

- # So let's apply the first difference and see if it helps

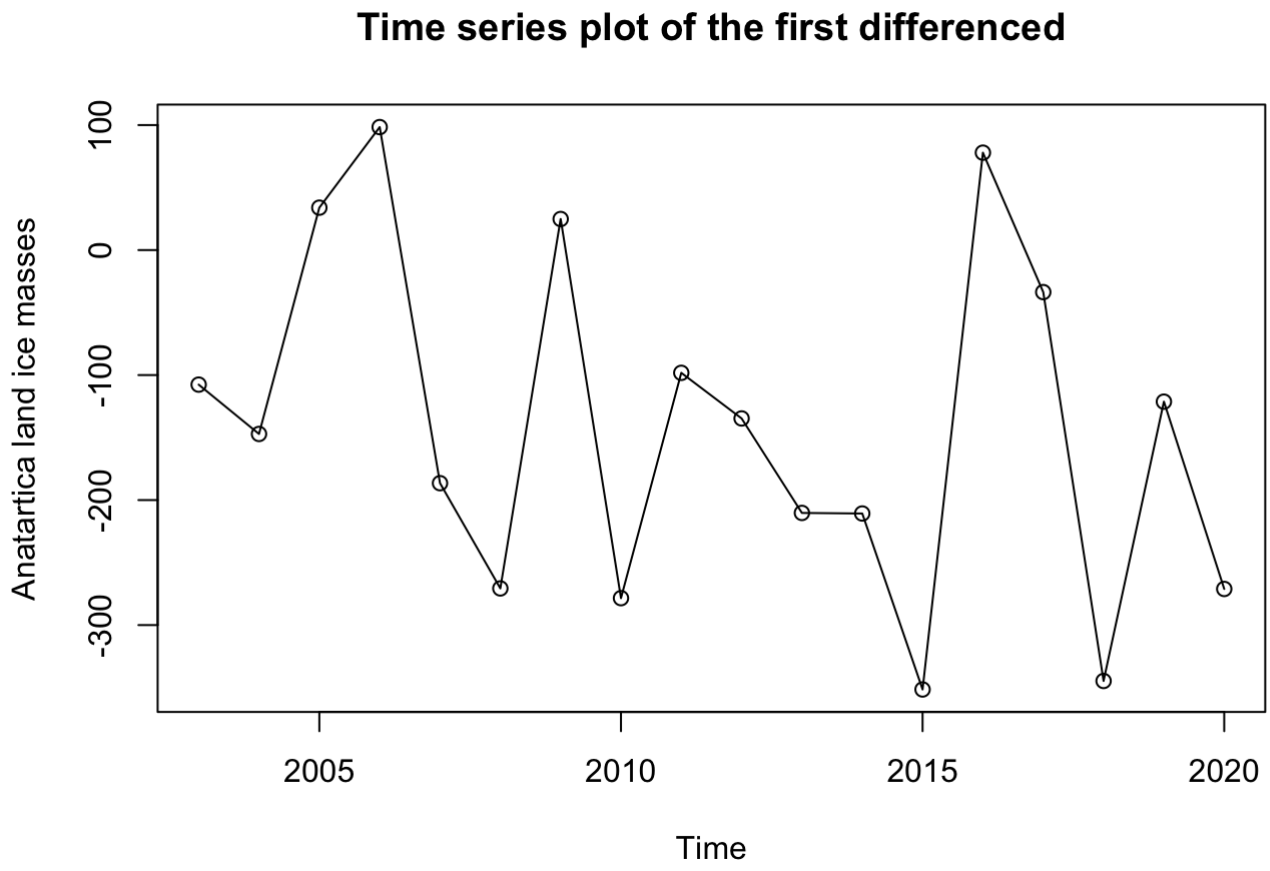
Differencing:

First Difference:

```
# So let's apply the first difference and see if it helps.

diff.data = diff(data1)

par(mfrow=c(1,1))
plot(diff.data,type='o',ylab='Anatartica land ice masses', main='Time series plot of
 the first differenced')
```




```
# Now, there is only changing variance in the series after taking the first differenc
e.
# Let's go on with the specification of the models.
adf.test(diff.data)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diff.data
## Dickey-Fuller = -2.5664, Lag order = 2, p-value = 0.3566
## alternative hypothesis: stationary
```

```
#pp test:
pp.test(diff.data)
```

```
##
##  Phillips-Perron Unit Root Test
##
## data:  diff.data
## Dickey-Fuller Z(alpha) = -17.99, Truncation lag parameter = 2, p-value
## = 0.04887
## alternative hypothesis: stationary
```

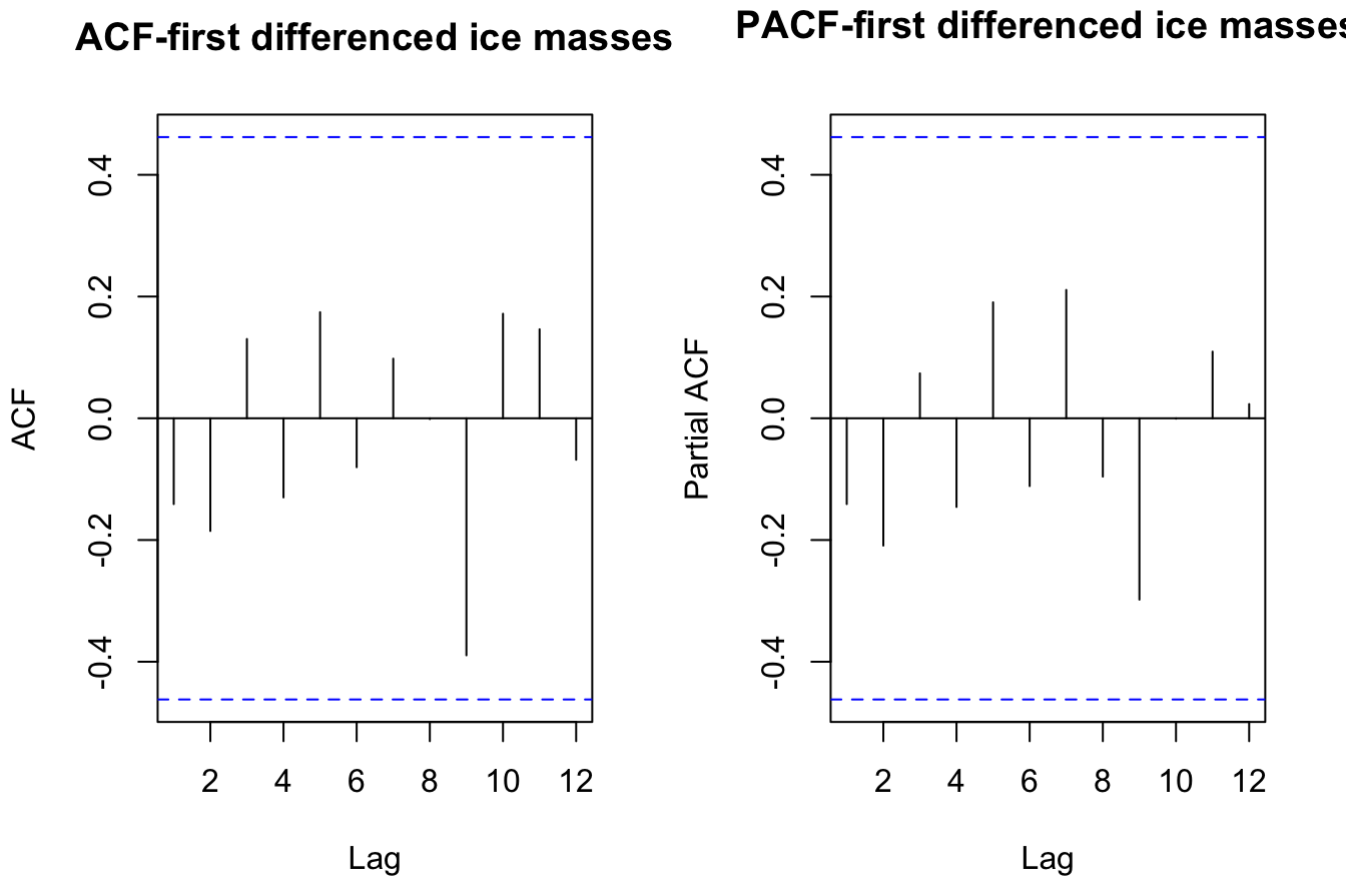
```
#kpss test:
kpss.test(diff.data)
```

```
## Warning in kpss.test(diff.data): p-value greater than printed p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  diff.data
## KPSS Level = 0.30558, Truncation lag parameter = 2, p-value = 0.1
```

```
par(mfrow=c(1,2))
# ACF and PACF plots:
acf(diff.data, main='ACF-first differenced ice masses')

pacf(diff.data, main ='PACF-first differenced ice masses')
```




```
par(mfrow=c(1,1))
```

Performing **ADF Test.** Output:

- The p-value is obtained is greater than significance level of 0.05 and the ADF statistic is higher than any of the critical values. Clearly, there is no reason to reject the null hypothesis. So, the time series is in fact non-stationary.
- Autocorrelation function (ACF), and Partial autocorrelation function (PACF): definition : ACF plot is a bar chart of coefficients of correlation between a time series and it lagged values. Simply stated: ACF explains how the present value of a given time series is correlated with the past (1-unit past, 2-unit past, …, n-unit past) values.

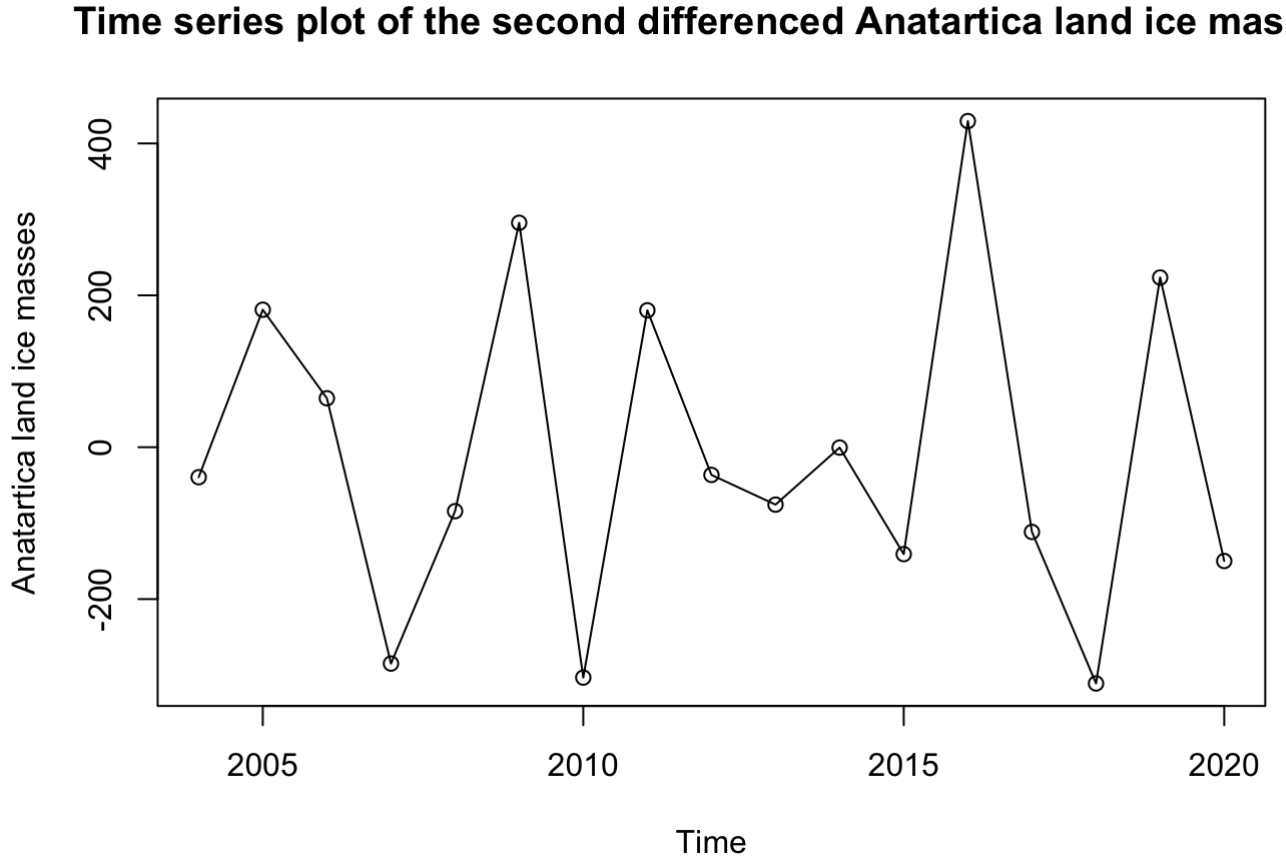
- PACF is the partial autocorrelation function that explains the partial correlation between the series and lags of itself.
- In ACF, we can see insignificant (spikes) lags are inside blue line), or no slowly decaying pattern.
- very high first correlation in PACF implies the existence of trend and non stationary.

So we will apply the second differencing.

Second Difference:

```
diff.data2 = diff(data, differences = 2)

par(mfrow=c(1,1))
plot(diff.data2,type='o',ylab='Anatartica land ice masses', main = "Time series plot
 of the second differenced Anatartica land ice masses")
```



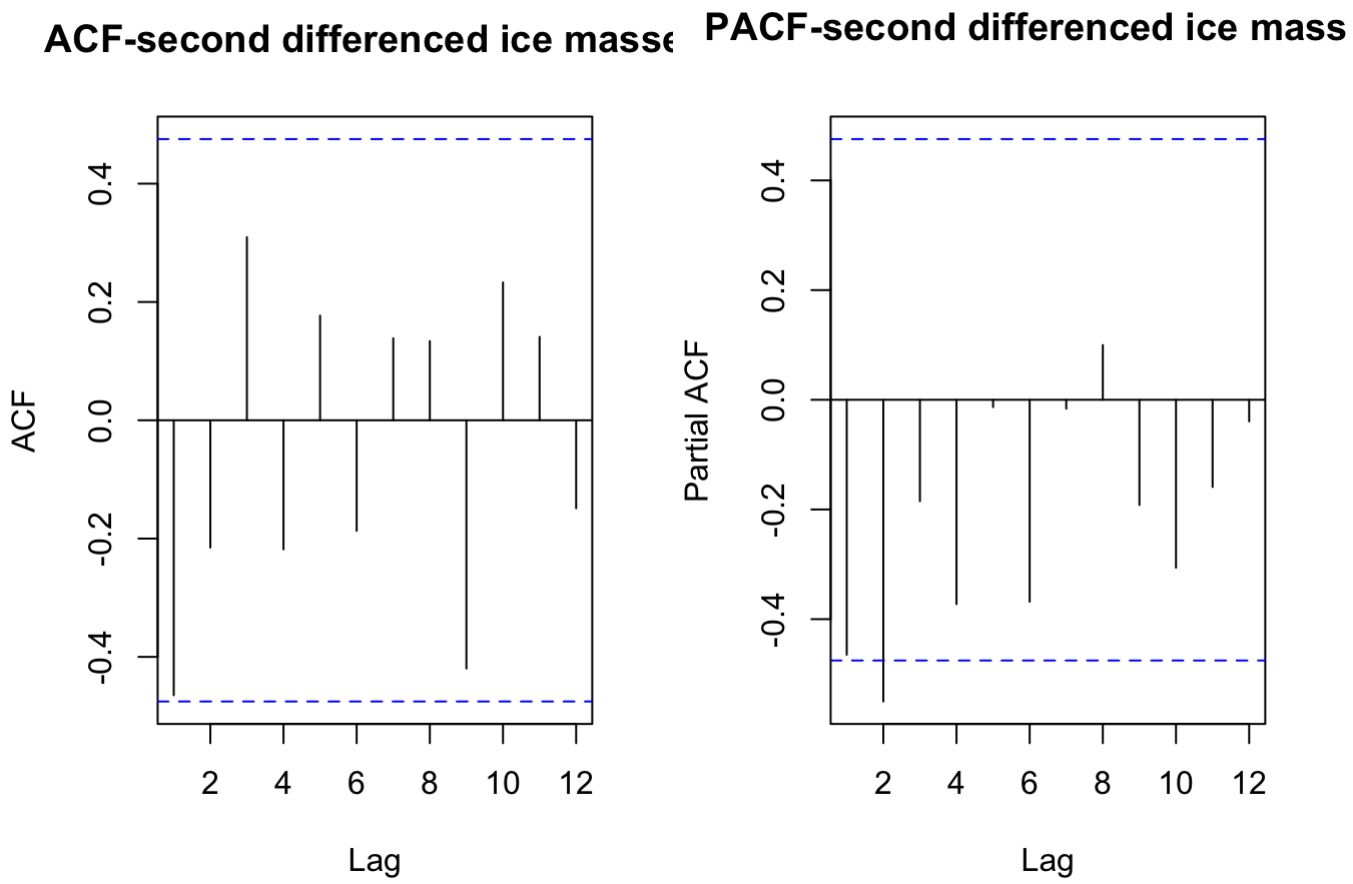
ADF Test:

- Performed ADF test: Output: In the results, ADF test concludes with the p-value of 0.07604 that the series is still non-stationary at 5% level of significance.
- But this p-value is close to the threshold 0.05.

```
adf.test(diff.data2)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diff.data2
## Dickey-Fuller = -3.438, Lag order = 2, p-value = 0.0725
## alternative hypothesis: stationary
```

- So we will look into ACF, PACF, PP test, and KPSS Test.
- Autocorrelation function (ACF), and Partial autocorrelation function (PACF): In ACF we can see lags are insignificant. No trend in the series implies that the series is stationary.
- By looking on ACF test and PACF , PP, KPSS test implies that the series is stationary, we can move for modeling.




```
##
##  Phillips-Perron Unit Root Test
##
## data:  diff.data2
## Dickey-Fuller Z(alpha) = -19.817, Truncation lag parameter = 2, p-value
## = 0.02604
## alternative hypothesis: stationary
```

```
## Warning in kpss.test(diff.data2): p-value greater than printed p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  diff.data2
## KPSS Level = 0.11554, Truncation lag parameter = 2, p-value = 0.1
```

- Output of PP test, p value in pp test is 0.02604, which is less than 0,05, implies stationary series.
- Output of KPss test, p value in pp test p-value = 0.1, which is greater than 0.05, implies stationary series.
- From ACF and PACF and PP test, KPSS, we conclude that the second difference of the series is stationary.
- Therefore, possible set of models:{p,d,q} by ACF and pcf we get,
- p= 2 , q= 2, if the pacf has 2 significant lags, q=2 when acf has 2 significant lags. ARIMA{(2,2,1), (1,2,1)}
- In pacf, 1st lag is touching the significant line(closely corelated), that is why we can take this value for possible set of models:

```
### **4. Model Fitting :**

- AutoRegressive Integrated Moving Average (ARIMA) models are among the most widely u
sed time series forecasting techniques: In an Autoregressive model, the forecasts cor
respond to a linear combination of past values of the variable.
So for selecting the best model, we need to find the possible set of models using EAC
F, Bic and ACF,PACF.

-   ACF and PACF : Because there is no significant lags seen in ACF, but first lag is
highly correlated (near to significance level) and we can consider it significant. an
d in PACF it can be see a second significant lag (highly correlated)

Possible set of ARMIA models:

ARIMA{(2,2,1),ARIMA(1,2,1)}

# EASF :

-   We then draw models from EACF tables which gives ARIMA{(1,2,2),(2,2,2),(1,2,),(2,
2,3)}.

-   Because of the size of the series we restrict the maximum number of AR & MA param
eters.

-   We put these argument to limit the orders p and q at 3.
    Otherwise, the eacf()function returns an error and displays nothing.


```r
#eacf(diff.data2)
eacf(diff.data2, ar.max = 3, ma.max = 3)
```
```

```
## AR/MA
##   0 1 2 3
## 0 o o o o
## 1 x x o o
## 2 o x o o
## 3 o o o o
```

```
# We put these argument to limit the orders p and q at 5. Otherwise, the eacf()
# function returns an error and displays nothing.
#possible set of models:{p,d,q}
#we can see Top let matrix
# ARIMA{(1,2,2),(2,2,2),(1,2,3),(2,2,3)
```

BIC :

We plot BIC table to get different ARIMA models :

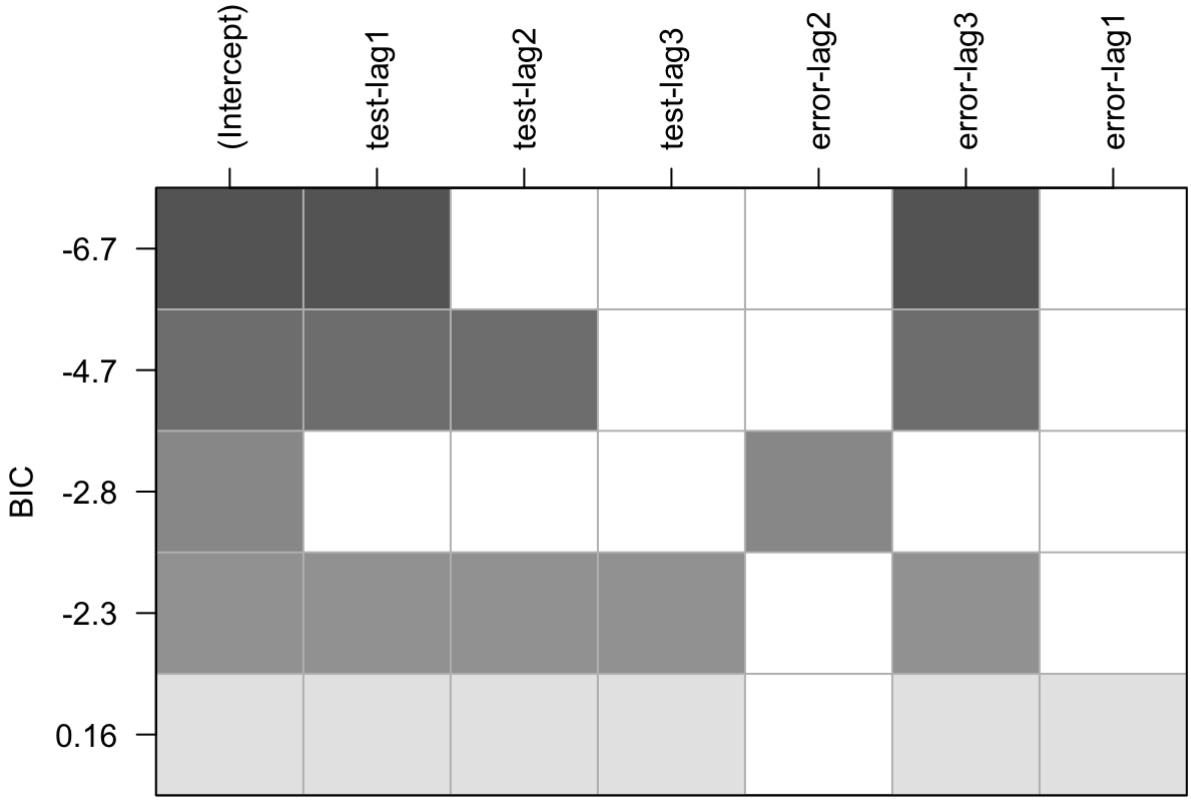
- In BIC table below shaded columns correspond to AR(1) and MA(3) coefficients.
- We put these argument to limit the orders p and q at 3. Otherwise, the eacf()function returns an error and displays nothing.
- ARIMA(1,2,3) and ARIMA(2,2,3) are the set of possible models we get through BIC table

```
# BIC
res=armasubsets(y=diff.data2,nar=3,nma=3,y.name='test',ar.method = 'yw')
```

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =
## force.in, : 1 linear dependencies found
```

```
## Reordering variables and trying again:
```

```
plot(res)
```




```
# {}
#possible sets for 1 best top row: ARIMA {1,2,3}
##or if we go for the second best BIC: we get p=1,2 and q=3.
# we have ARIMA{1,2,3}, {2,2,3}
```

7. Conclusion :

We analysed the yearly changes in Antarctica land ice mass in billion metric tons relative to the ice mass in 2001.We first converted the time series ,computing the correlation ,then transforming it by applying differencing (1st dirrerence and second difference). After every step we calculated ADF test, PP test, kpss sharpio-wilk test ACF and PACF plots to make sure series comes to be stationary in order to fit the model. In model fitting, We identified some possible set of models using EACF & BIC and estimated parameters using coefficient test.

Some information about ARIMA: ARIMA: Autoregressive Integrated Moving average First, made a non-stationary series stationary then use the ARMA model. ARIMA : P: Autoregressive order. D: The number of differences to make the non-stationary series stationary (i used 2 difference) Q: Moving average order:

After performing above mentioned models we have the possible set of models ARIMA(p,d,q):

ARIMA(2,2,1), ARIMA(1,2,1), ARIMA(1,2,2), ARIMA(2,2,2), ARIMA(1,2,3), ARIMA(2,2,3), ARIMA{1,2,3), ARIMA(1,2,3), ARIMA(2,2,3)

After fitting this models, we will find out the best model and can go for the forecasting.