DSC 425 Time Series Analysis Forecasting

Project: Sea Ice Extent

Exploratory Data Analysis

```
> seaice <- read_csv("Desktop/seaice.csv")
Rows: 26354 Columns: 7
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (2): Source Data, hemisphere
dbl (5): Year, Month, Day, Extent, Missing

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
> head(seaice)
# A tibble: 6 × 7
   Year Month   Day Extent Missing `Source Data`                                                                hemisphere
  <dbl> <dbl> <dbl>  <dbl>   <dbl> <chr>                                                                       <chr>
1  1978    10    26   10.2       0 ['ftp://sidads.colorado.edu/pub/DATASETS/nsidc0051_gsfc_nasateam_seaice/final-… north
2  1978    10    28   10.4       0 ['ftp://sidads.colorado.edu/pub/DATASETS/nsidc0051_gsfc_nasateam_seaice/final-… north
3  1978    10    30   10.6       0 ['ftp://sidads.colorado.edu/pub/DATASETS/nsidc0051_gsfc_nasateam_seaice/final-… north
4  1978    11     1   10.7       0 ['ftp://sidads.colorado.edu/pub/DATASETS/nsidc0051_gsfc_nasateam_seaice/final-… north
5  1978    11     3   10.8       0 ['ftp://sidads.colorado.edu/pub/DATASETS/nsidc0051_gsfc_nasateam_seaice/final-… north
6  1978    11     5   11.0       0 ['ftp://sidads.colorado.edu/pub/DATASETS/nsidc0051_gsfc_nasateam_seaice/final-… north
>
```

We removed the missing column and source data from our data, and after that combine the year, month, day into a single column and gave that column new name called date.

```
> seaice <- seaice[c(-5, -6)]
> seaice$Date <- as.Date(with(seaice, paste(Year, Month, Day, sep = '-')), "%Y-%m-%d")
> seaice <- seaice %>% select(-Year, -Month, -Day)
> head(seaice)
# A tibble: 6 × 3
  Extent hemisphere Date
   <dbl> <chr>      <date>
1   10.2 north      1978-10-26
2   10.4 north      1978-10-28
3   10.6 north      1978-10-30
4   10.7 north      1978-11-01
5   10.8 north      1978-11-03
6   11.0 north      1978-11-05
>
```

Moreover, we separate out the hemisphere into two parts i.e., North, and South.

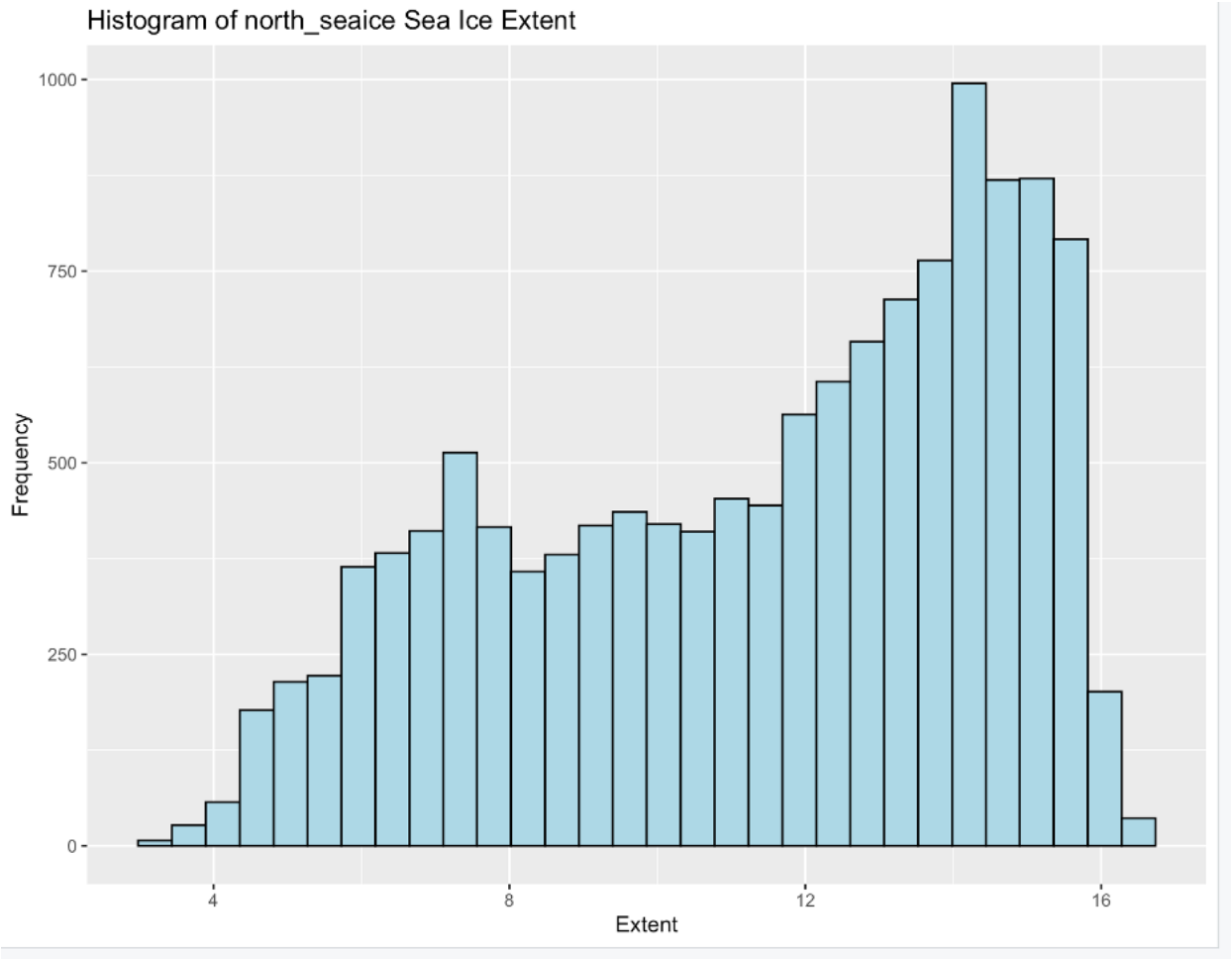
North Hemisphere:

```
> north_seaice <- filter(seaice, hemisphere == "north")
> head(north_seaice)
# A tibble: 6 × 3
  Extent hemisphere Date
   <dbl> <chr>      <date>
1   10.2 north      1978-10-26
2   10.4 north      1978-10-28
3   10.6 north      1978-10-30
4   10.7 north      1978-11-01
5   10.8 north      1978-11-03
6   11.0 north      1978-11-05
```

Plotting Histogram, QQ plot and Jarque Bera Test to check whether data is normally distributed.

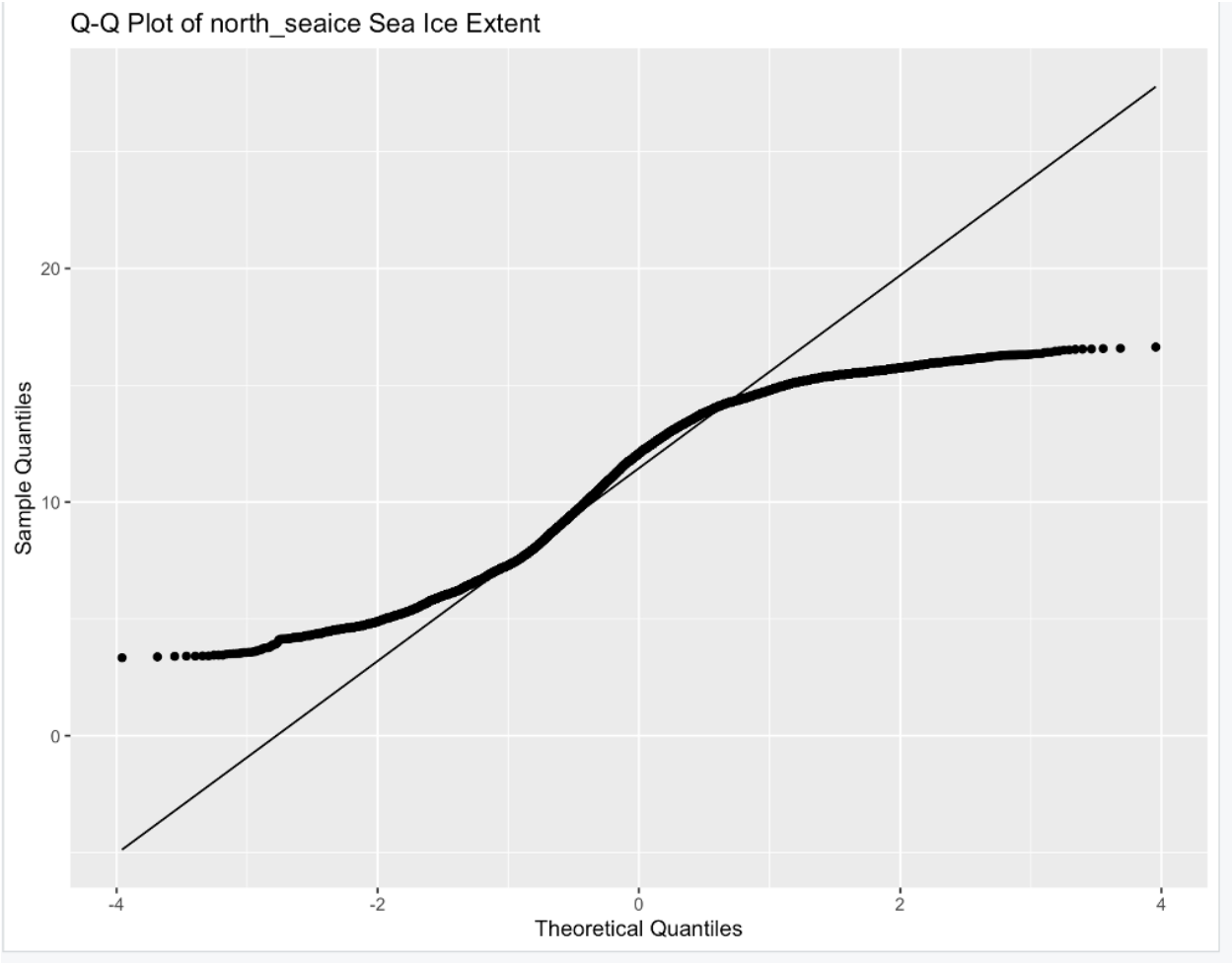
Histogram:

```
# Histogram north_seaice
ggplot(north_seaice, aes(x = Extent)) +
  geom_histogram( fill = "lightblue", color = "black") +
  labs(title = "Histogram of north_seaice Sea Ice Extent", x = "Extent", y = "Frequency")
```



QQ plot:

```
# Q-Q plot north_seaice
ggplot(north_seaice, aes(sample = Extent)) +
  geom_qq() +
  geom_qq_line() +
  labs(title = "Q-Q Plot of north_seaice Sea Ice Extent", x = "Theoretical Quantiles", y = "Sample Quantiles")
```



From the above plots, it seems that data is not normally distributed.

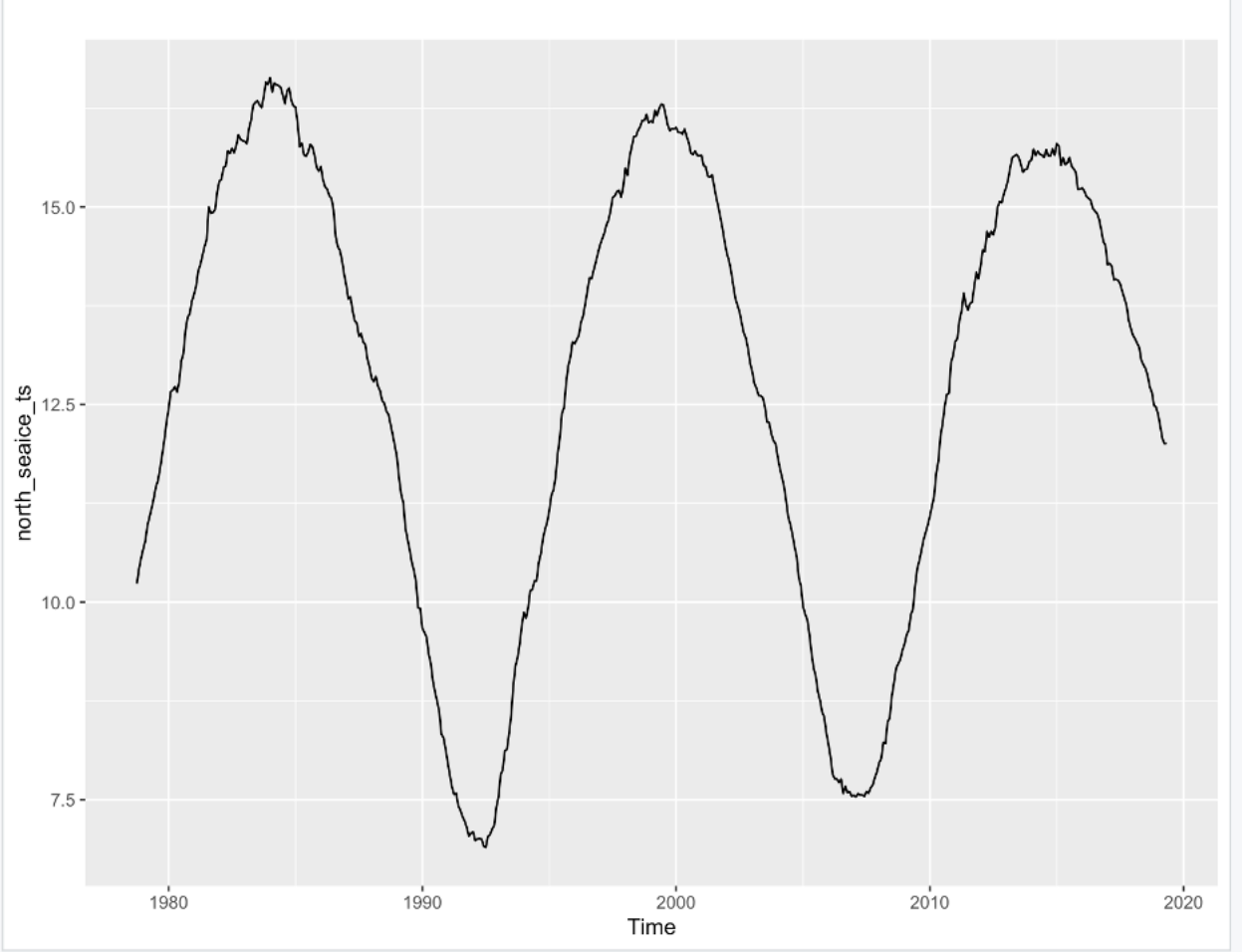
Jb Test:

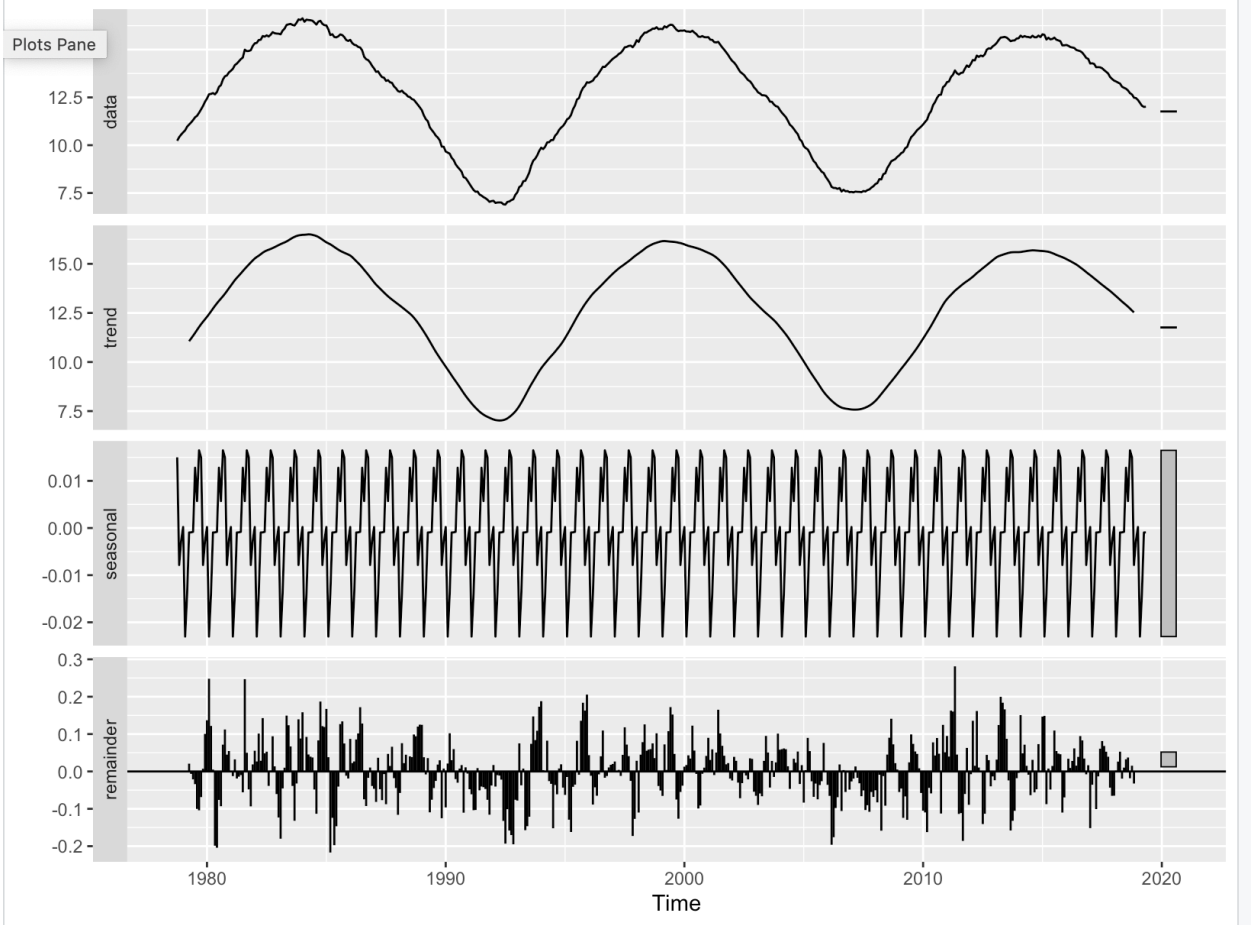
```
> jb_test <- jarque.bera.test(north_seaice$Extent)
>
> # Print the test results
> print(jb_test)

        Jarque Bera Test

data:  north_seaice$Extent
X-squared = 999.59, df = 2, p-value < 2.2e-16
```

Based on the Jarque-Bera test results you provided, with a test statistic of 999.59 and a p-value less than 2.2e-16, we can conclude that the data in north_seaice$Extent significantly deviates from a normal distribution. The extremely small p-value suggests strong evidence against the null hypothesis of normality.
Therefore, data does not follow a normal distribution based on this test.

Creating Time series:

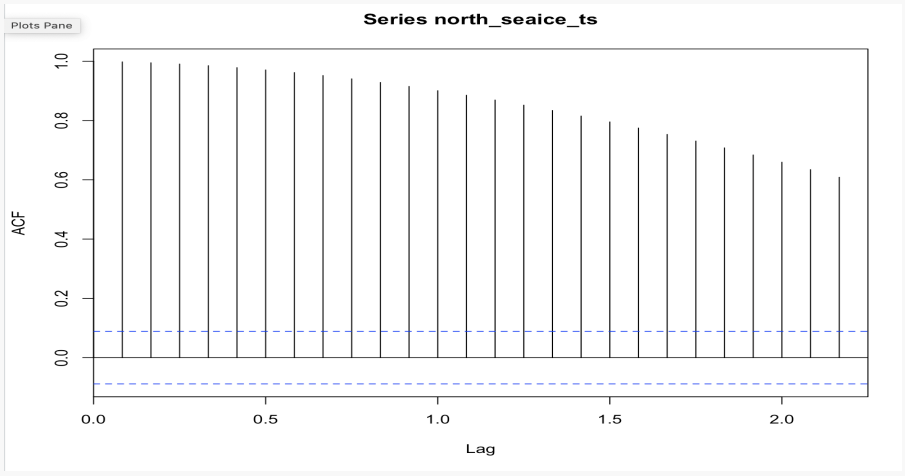
```
> north_seaice_ts = ts(north_seaice $Extent, start=c(1978, 10),end  = c(2019, 5), frequency=12)
> autoplot(north_seaice_ts)
>
```





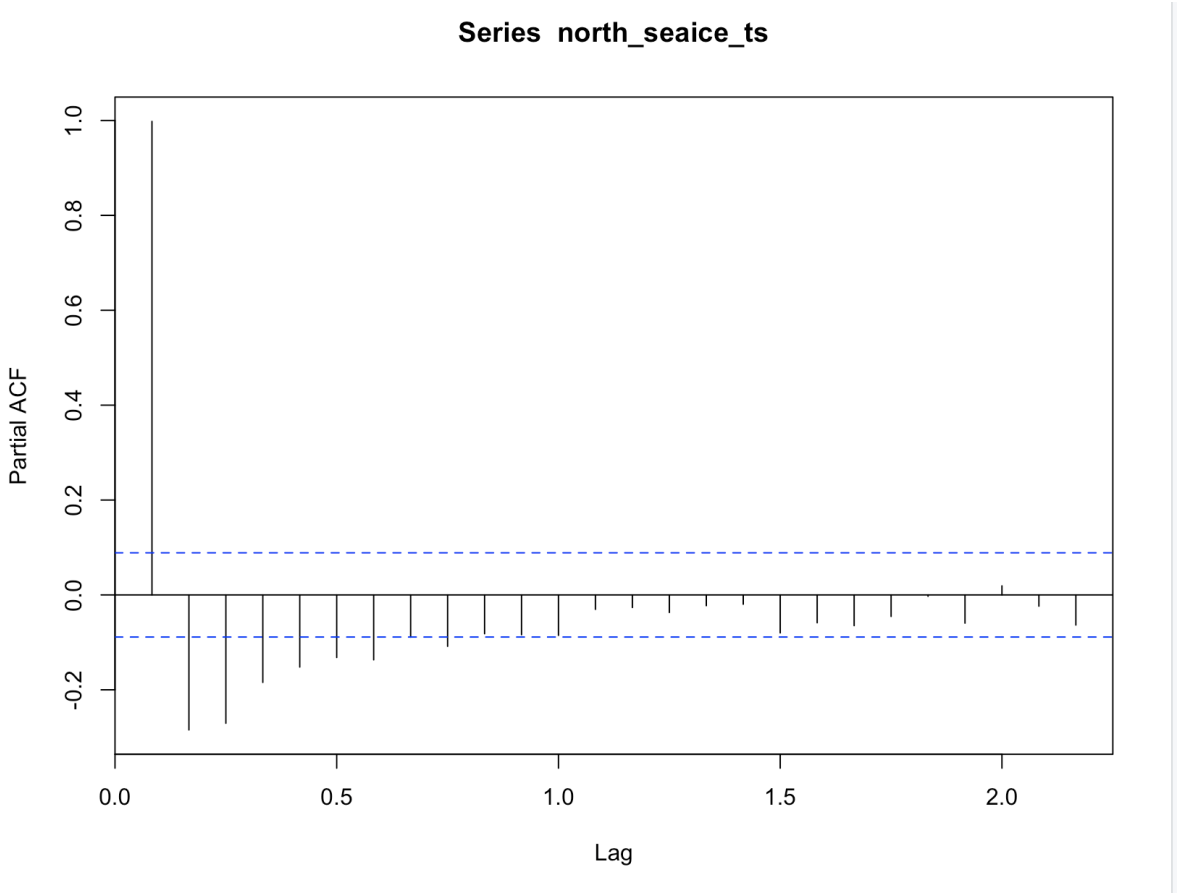
From the above plot we can see that there might be a seasonality present in data and seasonal term is tiny compared to the "remainder", and their appear to be periodic trend

acf Plot:



From the above plot we can see that there is a slow decrease, so it is non stationary.

Pacf Plot:



From the Pacf plot, there is a significant negative correlation at lag 2,3 and 4.

Ljung-Box test:

```
> Box.test(north_seaice_ts, type = "Ljung-Box")

        Box-Ljung test

data:  north_seaice_ts
X-squared = 489.1, df = 1, p-value < 2.2e-16
```

Based on these results, we can conclude that there is strong evidence against the null hypothesis of no autocorrelation in the **north_seaice_ts** data. The extremely small p-value (< 2.2e-16) suggests that there is significant autocorrelation present in the time series.

In summary, the Box-Ljung test indicates that the residuals of the **north_seaice_ts** data exhibit significant autocorrelation at different lags.

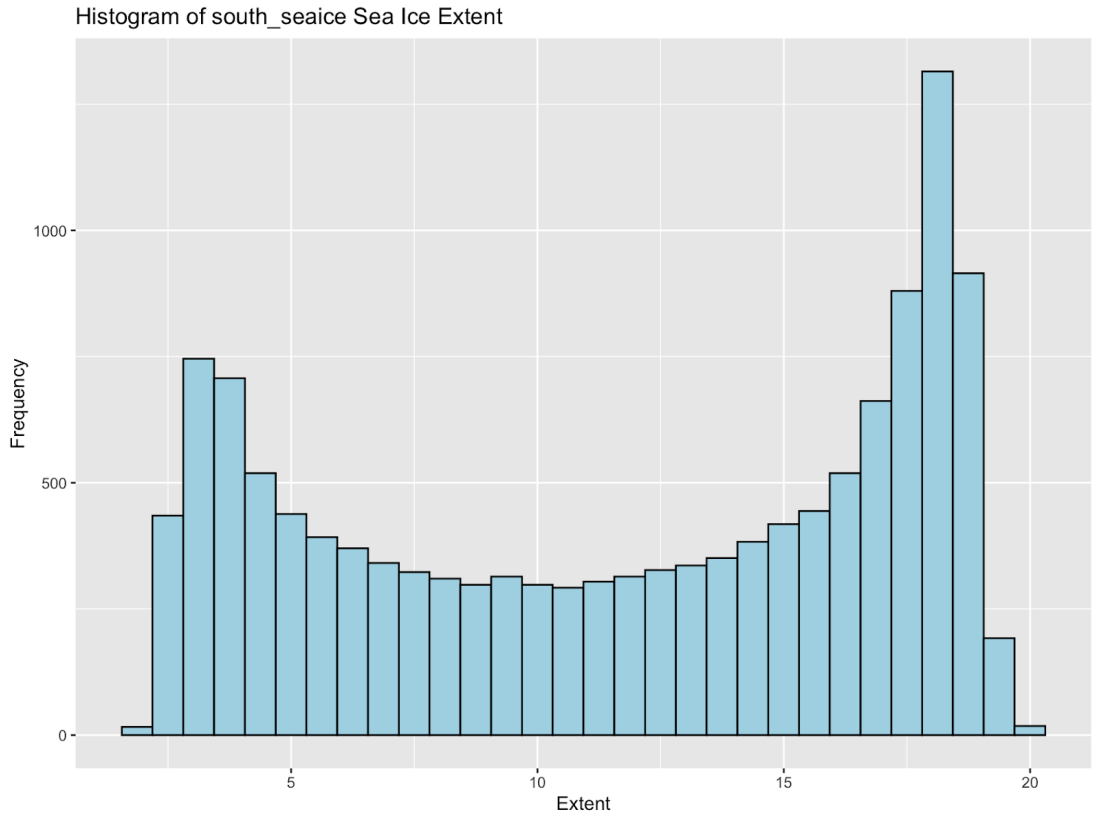
Dickey-Fuller and an KPSS unit root test:

```
> # Dickey-Fuller unit root test
> adf.test(north_seaice_ts)
Augmented Dickey-Fuller Test
alternative: stationary

Type 1: no drift no trend
     lag       ADF p.value
[1,]   0  0.406024   0.761
[2,]   1  0.000257   0.644
[3,]   2 -0.295669   0.559
[4,]   3 -0.470461   0.509
[5,]   4 -0.608167   0.461
[6,]   5 -0.821923   0.385
Type 2: with drift no trend
     lag    ADF p.value
[1,]   0 -0.858  0.7517
[2,]   1 -1.088  0.6702
[3,]   2 -1.615  0.4819
[4,]   3 -2.033  0.3149
[5,]   4 -2.386  0.1736
[6,]   5 -2.839  0.0553
Type 3: with drift and trend
     lag    ADF p.value
[1,]   0 -0.878   0.955
[2,]   1 -1.099   0.923
[3,]   2 -1.619   0.739
[4,]   3 -2.035   0.562
[5,]   4 -2.388   0.412
[6,]   5 -2.838   0.223
----
Note: in fact, p.value = 0.01 means p.value <= 0.01
>
> # KPSS unit root test
> kpss.test(north_seaice_ts)
KPSS Unit Root Test
alternative: nonstationary

Type 1: no drift no trend
 lag  stat p.value
   5 0.357     0.1
-----
 Type 2: with drift no trend
 lag  stat p.value
   5 0.334     0.1
-----
 Type 1: with drift and trend
 lag  stat p.value
   5 0.309    0.01
-----------
Note: p.value = 0.01 means p.value <= 0.01
    : p.value = 0.10 means p.value >= 0.10
```

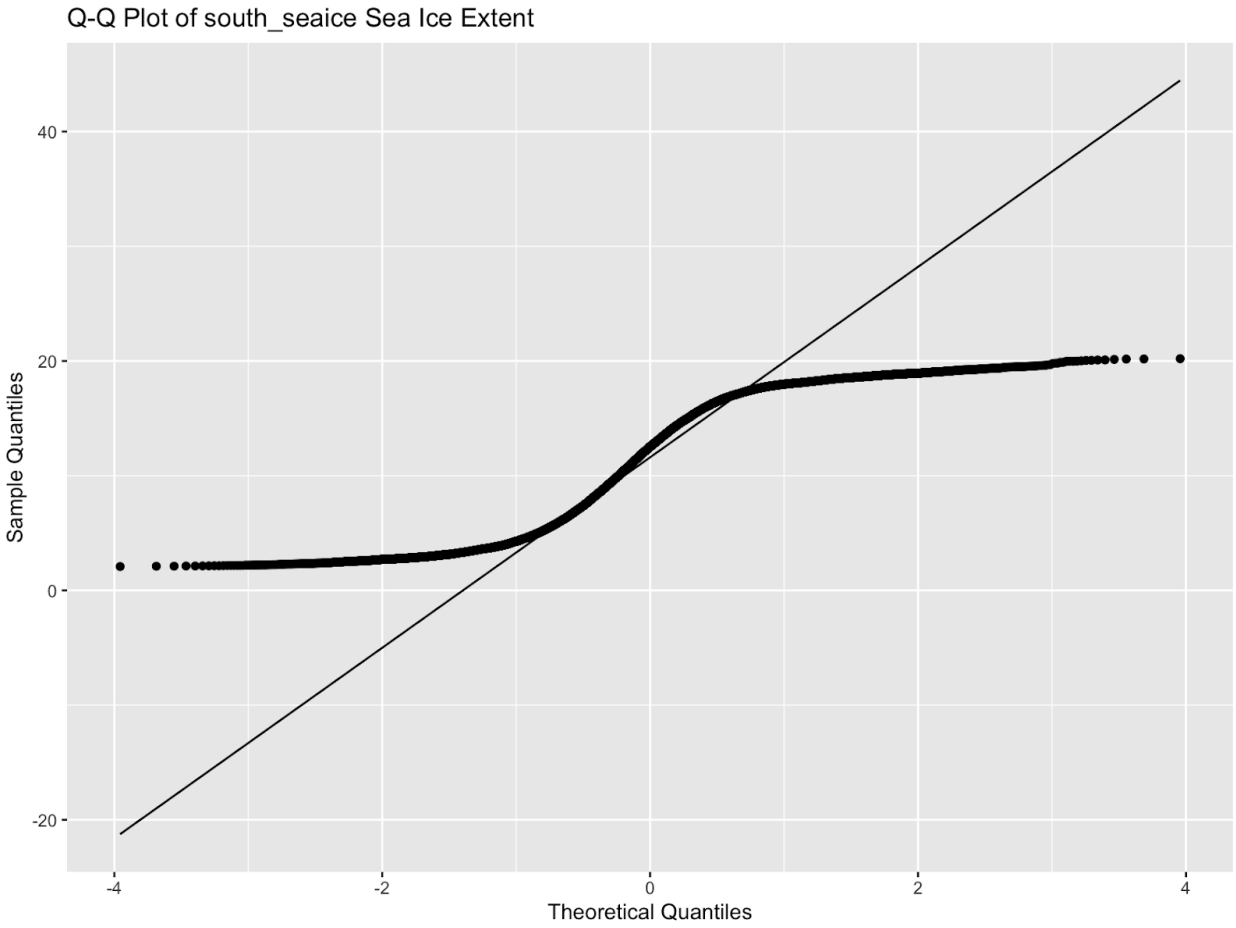
Based on the results of the Dickey-Fuller test, the p-values for all three types (no drift no trend, with drift no trend, and with drift and trend) are greater than 0.05. This suggests that we fail to reject the null hypothesis of the Dickey-Fuller test, indicating that the series is non-stationary. Similarly, the KPSS unit root test results show that the p-values for all three types (no drift no trend, with drift no trend, and with drift and trend) are greater than 0.01. This implies that we fail to reject the null hypothesis of the KPSS test, indicating non-stationarity in the series.
In summary, both the Dickey-Fuller and KPSS tests indicate that the series is non-stationary, meaning it does not exhibit a constant mean and variance over time.

South Hemisphere:

Histogram of south_seaice Sea Ice Extent:



South Hemisphere QQ Plot:J



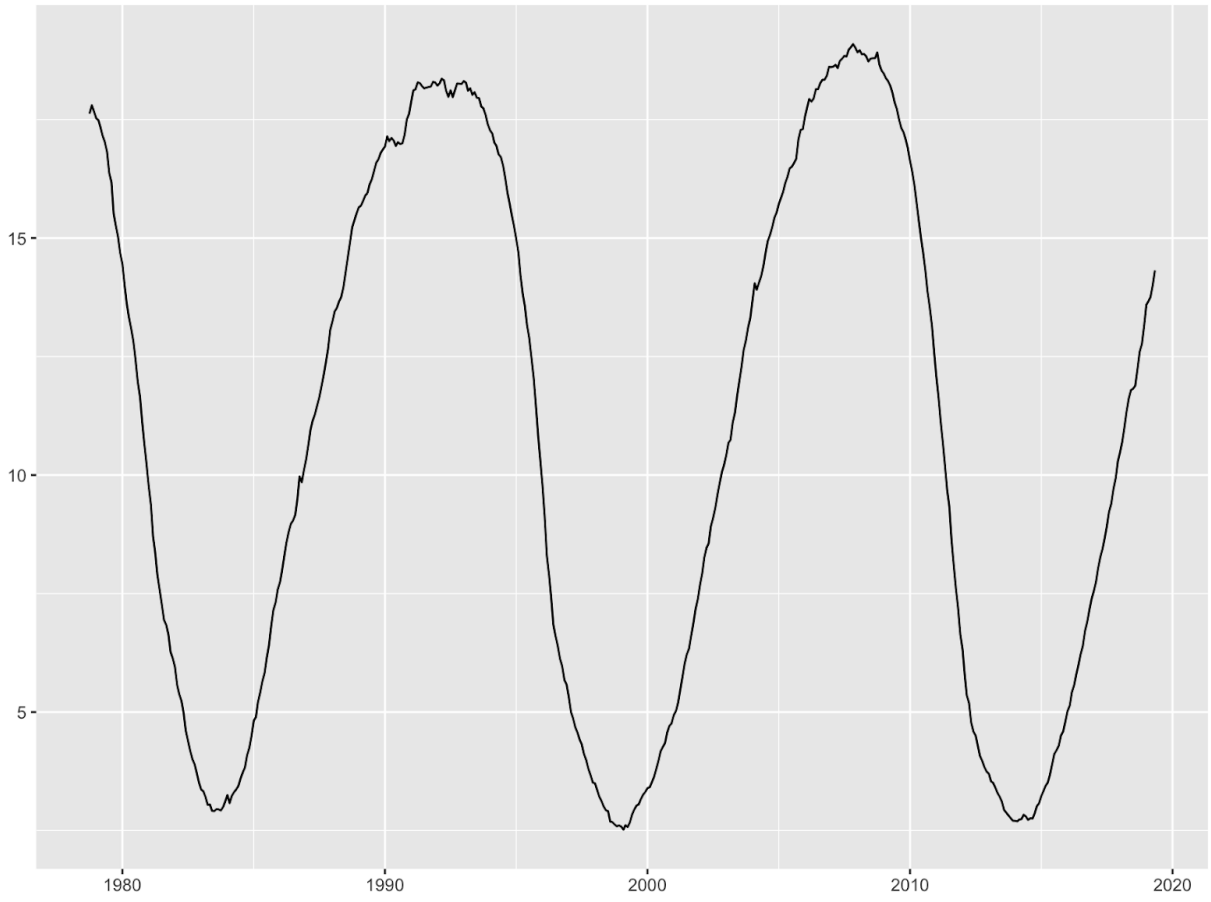
=> Based on the above plot we can interpret that south data is not normally distributed.

JB Test:

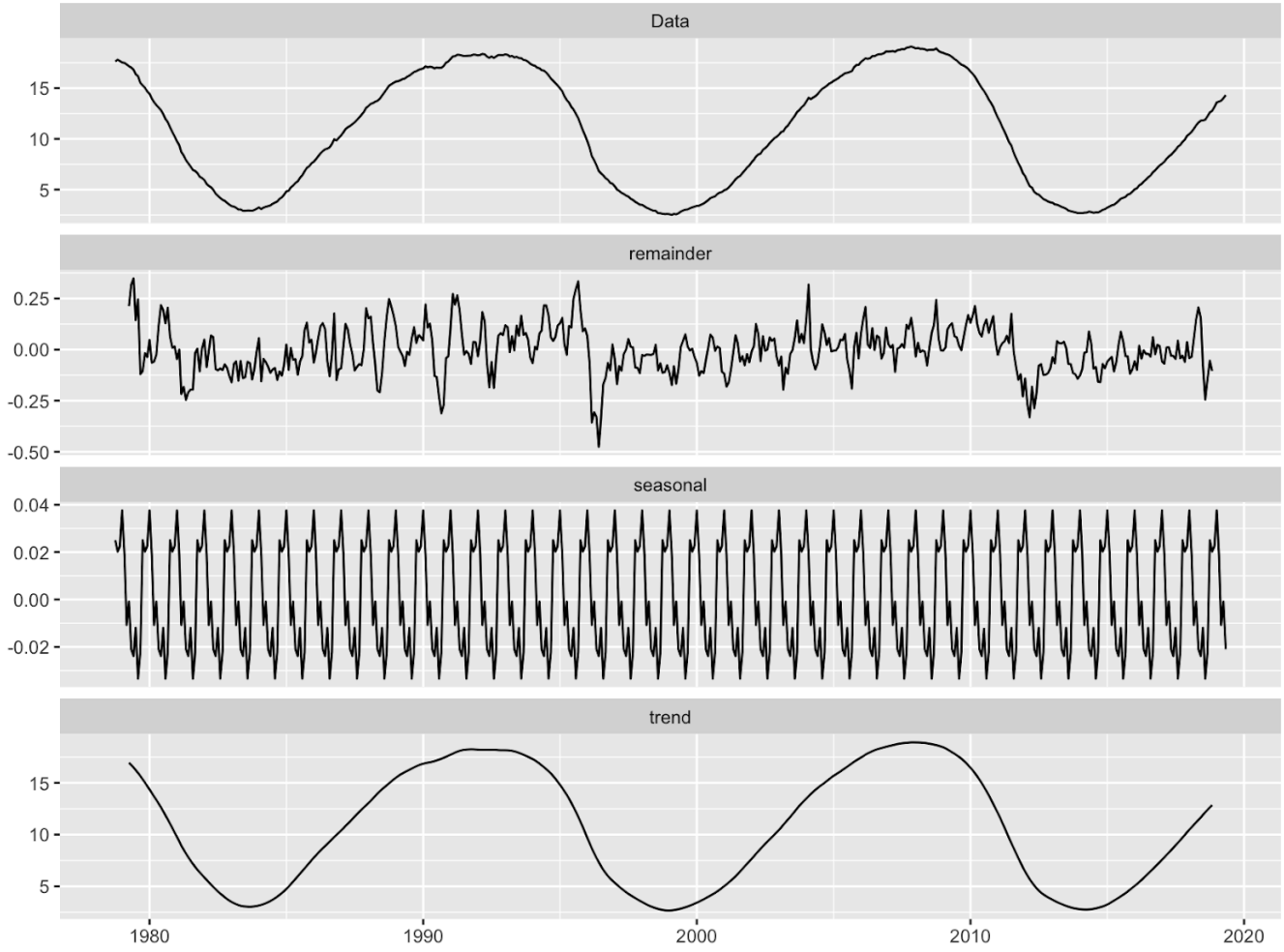
```
> jb_test <- jarque.bera.test(south_seaice$Extent)
> # Print the test results
> print(jb_test)

        Jarque Bera Test

data:  south_seaice$Extent
X-squared = 1285.9, df = 2, p-value < 2.2e-16

>
```

=> Based on the above results of JB test, we see that degree of freedom is 2, X-squared value is 1285.9 and p-values is less than significant value(0.05). The low p-value tells us that we can reject null hypothesis. This suggests us that our data is not normally distributed.

Creating Time series:

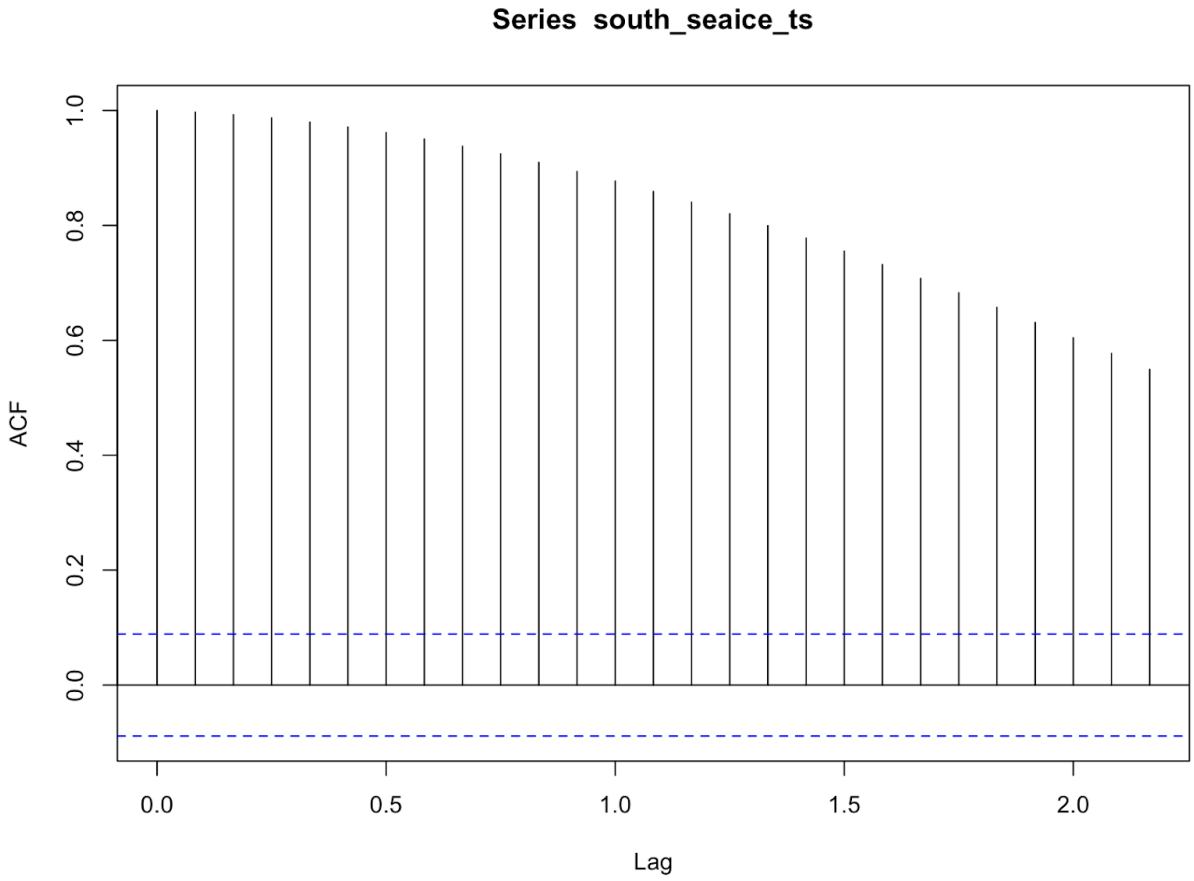


Decompose:



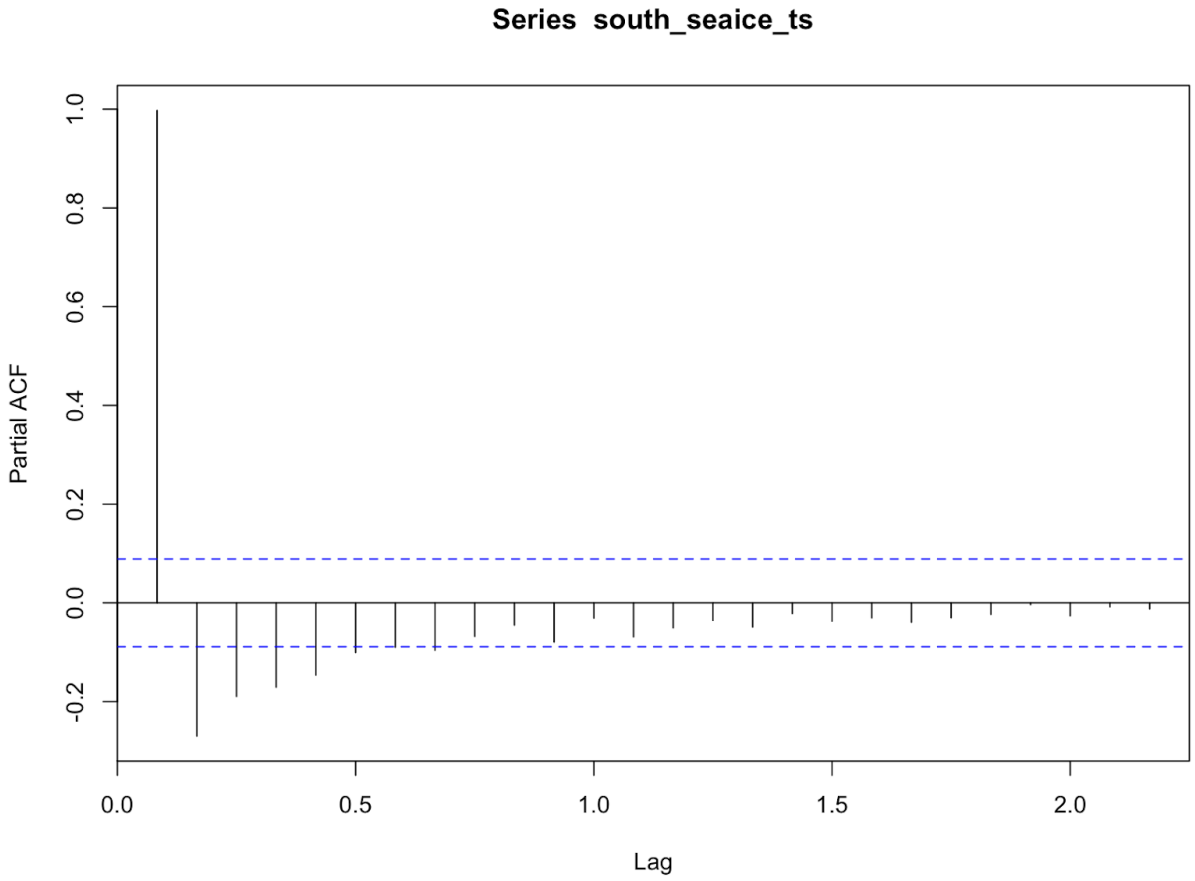
The above plot tell us about trend, seasonality and remainder. The trend component tells us that the some percentage of decline in sea ice extent. The seasonal component tells us that there is some cyclic pattern in sea ice extent indicating that there is some seasonality. The remainder plot tell us that there are random fluctuations in the sea ice extent.

ACF:



Based on the ACF we can see that there is slow decay and the series is non-stationary.

PCAF:



=> Based on the above PACF plot we can see that there is significant negative correlation at lag 2,3 and 4

Ljung Test:

```
> Box.test(south_seaice_ts,type='Ljung')

        Box-Ljung test

data:  south_seaice_ts
X-squared = 488.23, df = 1, p-value < 2.2e-16
```

Based on the above results of Ljung test, we see that degree of freedom is 1, X-squared value is 488.23 and p-values is less than significant value(0.05). The low p-value tells us that we can reject null hypothesis.

Dickey-fuller test:

```
> # Dickey-Fuller unit root test
> adf.test(south_seaice_ts)

        Augmented Dickey-Fuller Test

data:  south_seaice_ts
Dickey-Fuller = -4.8777, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary
```

Based on the above results of Dickey-Fuller test, we see that lag order is 7, Dickey-Fuller value is -4.8777 and p-values is less than significant value(0.05). The low p-value tells us that we can reject null hypothesis. The results shows that the series is stationary, however there is a discrepancy when we look at ACF and PACF plots we can tell that the series is non-stationary. We need further investigation.

Kpss test:

```
> kpss.test(south_seaice_ts)

        KPSS Test for Level Stationarity

data:  south_seaice_ts
KPSS Level = 0.37409, Truncation lag parameter = 5, p-value = 0.08832

>
```

Based on the KPSS test results, we see that KPSS level statistic is 0.37409, the truncation lag parameter used is 5. The p-value is 0.08832 which is greater than significant value. Therefore we fail to reject null hypothesis of level stationary. This mean that time series is stationary at the given level. But in contrast the p-value is greater than significant value which means that we need more investigation.

Further analysis to be done:

Based on the feedback received from our professor, we made some changes to our analysis approach and decided to focus on the North and South hemispheres separately.

After examining the results obtained thus far, we have concluded that the time series data for the North hemisphere is non-stationary. In order to address this issue, we will apply differencing to the series.

By taking the difference between consecutive observations, we aim to transform the data into a stationary series.
This process involves subtracting each data point from its previous one.
Once we have obtained the differenced series, we will proceed with further analysis and evaluation. This approach allows us to investigate the patterns and trends within the data while accounting for the non-stationarity observed initially.