ENV 790.30 - Time Series Analysis for Energy Data | Spring 2024

Assignment 4 - Due date 02/12/24

Julia Kagiliery

Directions

You should open the .rmd file corresponding to this assignment on RStudio. The file is available on our class repository on Github. And to do so you will need to fork our repository and link it to your RStudio.

Once you have the file open on your local machine the first thing you will do is rename the file such that it includes your first and last name (e.g., "LuanaLima_TSA_A04_Sp23.Rmd"). Then change "Student Name" on line 4 with your name.

Then you will start working through the assignment by **creating code and output** that answer each question. Be sure to use this assignment document. Your report should contain the answer to each question and any plots/tables you obtained (when applicable).

When you have completed the assignment, **Knit** the text and code into a single PDF file. Submit this pdf using Sakai.

R packages needed for this assignment: "xlsx" or "readxl", "ggplot2", "forecast","tseries", and "Kendall". Install these packages, if you haven't done yet. Do not forget to load them before running your script, since they are NOT default packages.\

```
library(readxl)
library(ggplot2)
library(forecast)
```

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

```
library(tseries)
library(Kendall)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(cowplot)
```

Questions

Consider the same data you used for A3 from the spreadsheet "Table_10.1_Renewable_Energy_Production_and_Consumption The data comes from the US Energy Information and Administration and corresponds to the January 2021 Monthly Energy Review. For this assignment you will work only with the column "Total Renewable Energy Production".

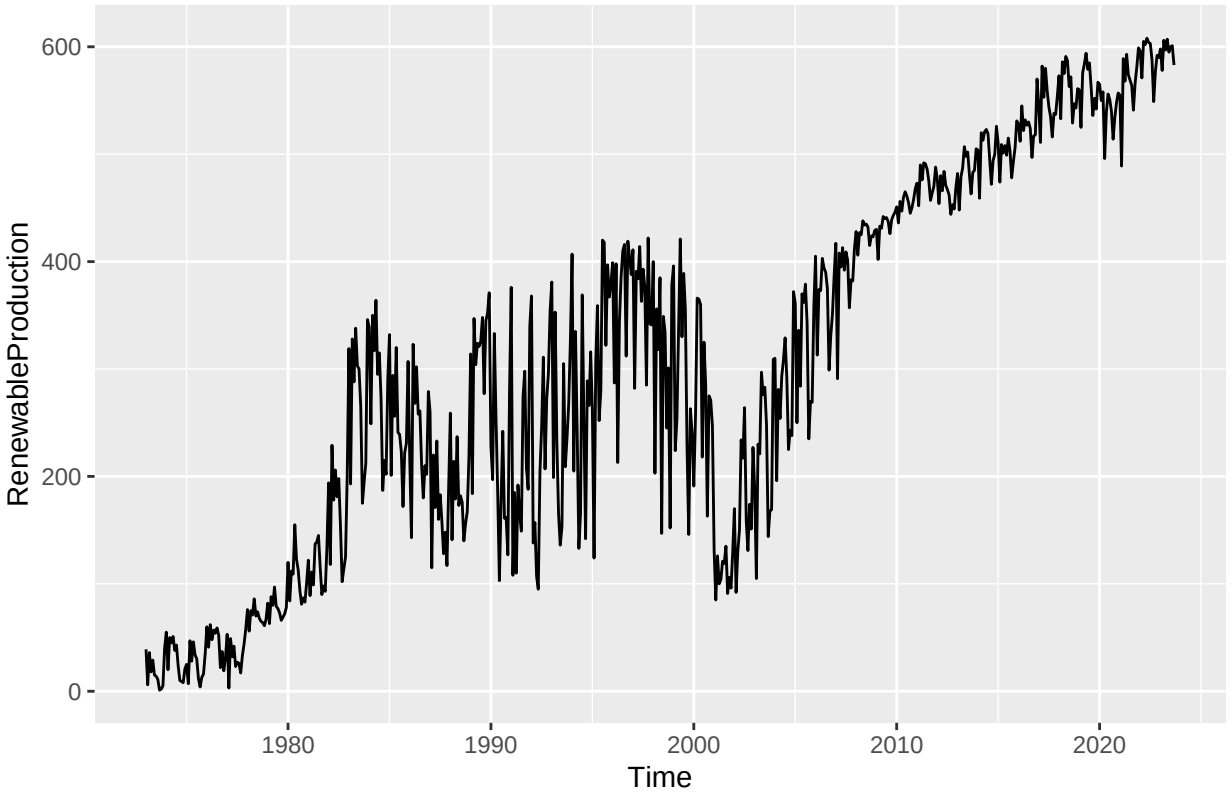
```
#Importing data set - using readxl package
RenewableProduction <- read_excel("~/Julia_Kagiliery_TSA_Sp24/Data/Table_10.1_Renewable_Energy_Producti
    skip = 10)
```

```
RenewableProductionDates <- RenewableProduction[-1,1]
```

```
RenewableProduction <- RenewableProduction[-1,] |>
  select(`Total Renewable Energy Production`)
```

```
year1 <- 1973
month1<- 1
RenewableProduction <- ts(RenewableProduction, start = c(year1, month1),
     frequency = 12)
```

```
autoplot(RenewableProduction)
```



Stochastic Trend and Stationarity Tests

Q1

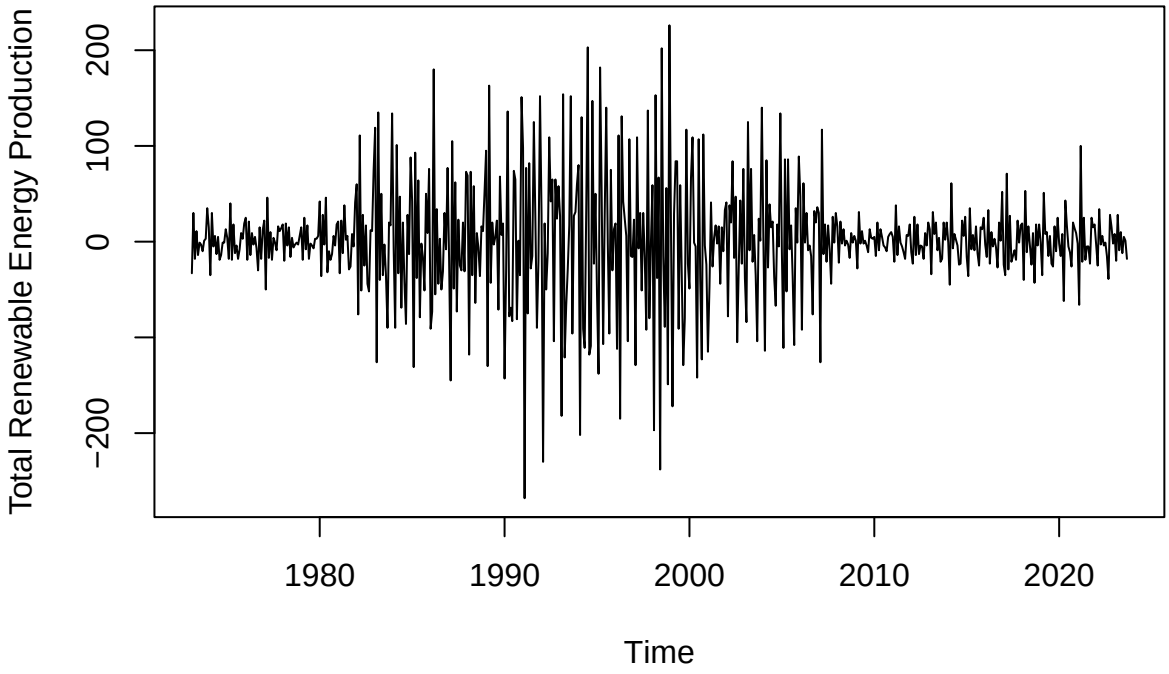
Difference the "Total Renewable Energy Production" series using function diff(). Function diff() is from package base and take three main arguments: * x vector containing values to be differenced; * *lag* integer

indicating with lag to use; * *differences* integer indicating how many times series should be differenced.

Try differencing at lag 1 only once, i.e., make `lag=1` and `differences=1`. Plot the differenced series Do the series still seem to have trend?

```
RenewableProductiondiff <- diff(RenewableProduction, 1, 1)
```

```
RenewableProductiondiffPlot <- plot(RenewableProductiondiff)
```



Q2

Copy and paste part of your code for A3 where you run the regression for Total Renewable Energy Production and subtract that from the orinal series. This should be the code for Q3 and Q4. make sure you use the same name for you time series object that you had in A3.

```
t <- c(1:609)
RenewableTrend = lm(RenewableProduction ~ t)
Rbeta0 = as.numeric(RenewableTrend$coefficients[1])
Rbeta1 = as.numeric(RenewableTrend$coefficients[2])
DataRenewableTrend <- Rbeta0 + (Rbeta1 * t)
```

```
RenewableProduction <- as.numeric(RenewableProduction)
DataRenewableTrend <- as.numeric(DataRenewableTrend)
DetrendedRenewable <- RenewableProduction - DataRenewableTrend
DetrendedRenewable <-
  ts(DetrendedRenewable,
     start = c(year1, month1),
     frequency = 12)
```

```
DetrendedRenewablePlot <- DetrendedRenewable |>
  autoplot(color = "blue") +
  ylab("Total Renewabke Energy Production [Trillion BTU]") +
  xlab("Year")
```

Q3

Now let's compare the differenced series with the detrended series you calculated on A3. In other words, for the "Total Renewable Energy Production" compare the differenced series from Q1 with the series you detrended in Q2 using linear regression.

Using autoplot() + autolayer() create a plot that shows the three series together. Make sure your plot has a legend. The easiest way to do it is by adding the `series=` argument to each autoplot and autolayer function. Look at the key for A03 for an example.

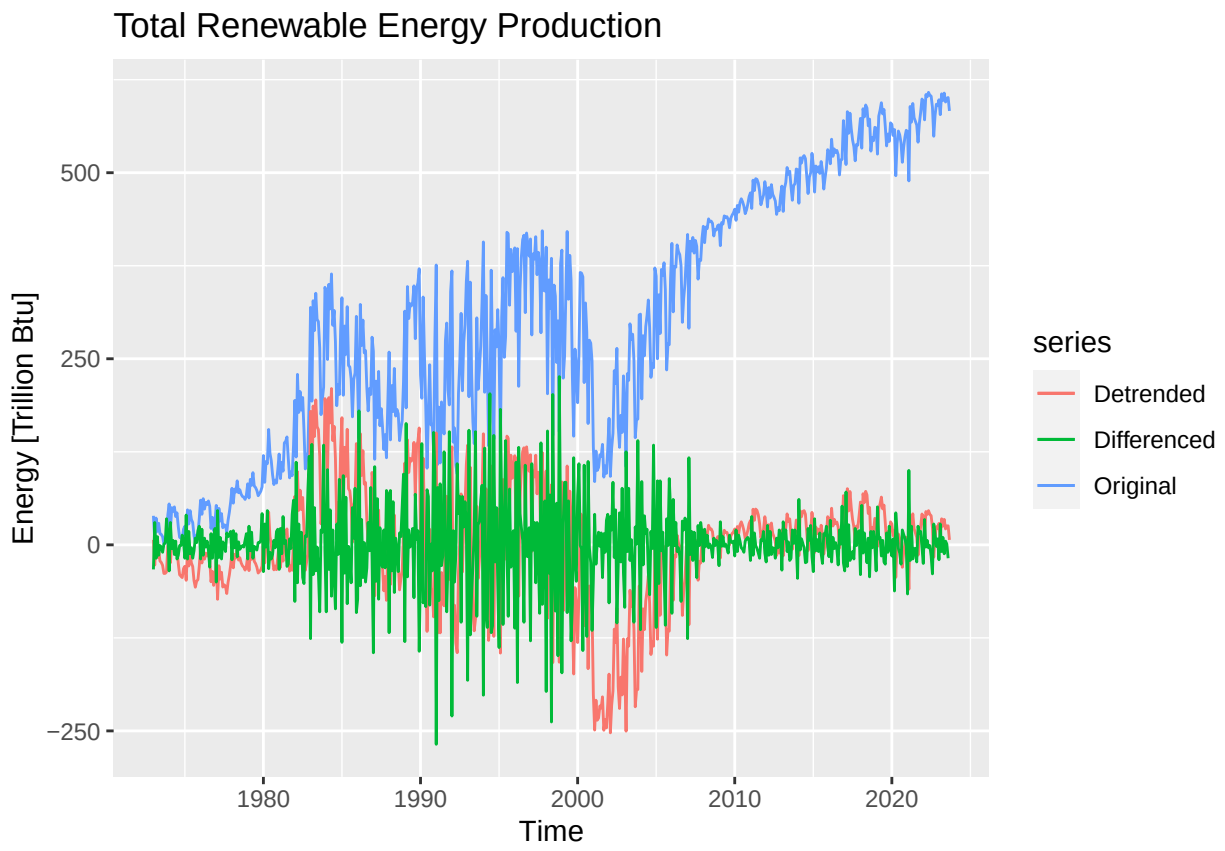
Three series: as is, detrended, differenced

```
DetrendedRenewable <-
  ts(DetrendedRenewable,
     start = c(year1, month1),
     frequency = 12)

RenewableProduction <-
  ts(RenewableProduction,
     start = c(year1, month1),
     frequency = 12)

RenewableProductiondiff <-
  ts(RenewableProductiondiff,
     start = c(year1, month1),
     frequency = 12)

autoplot(RenewableProduction,series="Original") +
autolayer(DetrendedRenewable,series="Detrended") +
  autolayer(RenewableProductiondiff,series="Differenced") +
ylab("Energy [Trillion Btu]") +
ggtitle("Total Renewable Energy Production")
```

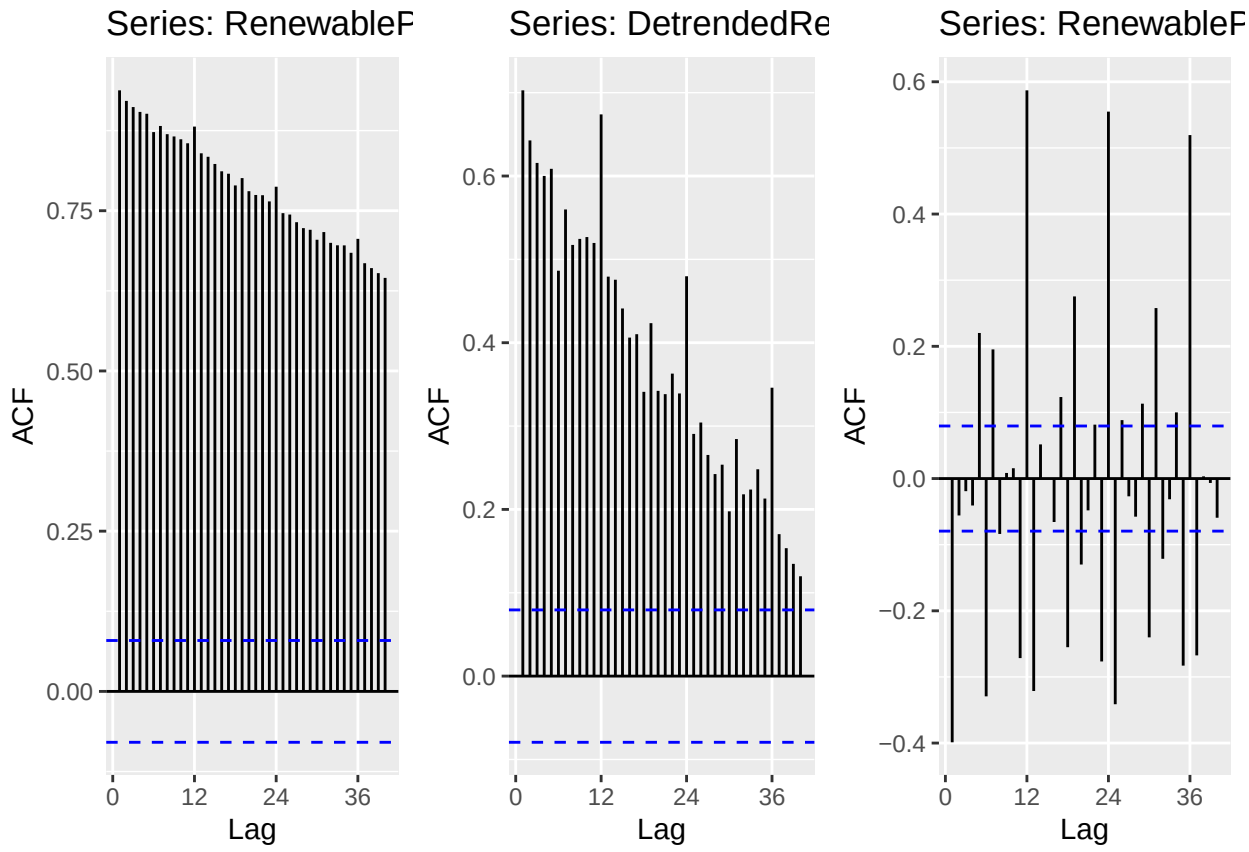


Q4

Plot the ACF for the three series and compare the plots. Add the argument `ylim=c(-0.5,1)` to the autoplot() or Acf() function - whichever you are using to generate the plots - to make sure all three y axis have the same limits. Which method do you think was more efficient in eliminating the trend? The linear regression or differencing?

```
P2 <-autoplot(Acf(DetrendedRenewable, lag.max = 40, ylim=c(-0.5,1), plot = FALSE))
P1 <- autoplot(Acf(RenewableProduction, lag.max = 40, ylim=c(-0.5,1), plot = FALSE))
P3 <- autoplot(Acf(RenewableProductiondiff, lag.max = 40, ylim=c(-0.5,1), plot = FALSE))
```

```
plot_grid(P1, P2, P3, nrow = 1)
```



The differenced ACF seems to look promising as is has relatively few spikes beyond the blue dashed significance lines. The Detrended ACF also looks better as it is downward sloping with lower values and some regular spikes at 12-month intervals.

Q5

Compute the Seasonal Mann-Kendall and ADF Test for the original "Total Renewable Energy Production" series. Ask R to print the results. Interpret the results for both test. What is the conclusion from the Seasonal Mann Kendall test? What's the conclusion for the ADF test? Do they match what you observed in Q3 plot? Recall that having a unit root means the series has a stochastic trend. And when a series has stochastic trend we need to use a different procedure to remove the trend.

```
print("Results for ADF test/n")
```

```
## [1] "Results for ADF test/n"
```

```
print(adf.test(RenewableProduction,alternative = "stationary"))
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  RenewableProduction
## Dickey-Fuller = -3.1076, Lag order = 8, p-value = 0.1095
## alternative hypothesis: stationary
```

The p-value for this ADF is 0.1095 hence we fail to reject the null hypothesis that the data is not stationary. In this case, it seems like our data is not stationary.

```r
SMKtest <- SeasonalMannKendall(RenewableProduction)
print("Results for Seasonal Mann Kendall /n")
```

```
## [1] "Results for Seasonal Mann Kendall /n"
```

```r
print(summary(SMKtest))
```

```
## Score =  11865 , Var(Score) = 179299
## denominator =  15149.5
## tau = 0.783, 2-sided pvalue =< 2.22e-16
## NULL
```

The p-value for this test is ~0 hence we reject the null hypothesis and belive that there is a trend.

This looks about right compared to Q3; the original data is definitely trended and seasonal.

Q6

Aggregate the original "Total Renewable Energy Production" series by year. You can use the same procedure we used in class. Store series in a matrix where rows represent months and columns represent years. And then take the columns mean using function colMeans(). Recall the goal is the remove the seasonal variation from the series to check for trend. Convert the accumulates yearly series into a time series object and plot the series using autoplot().

```r
RenewableProduction_year <- as.integer(format(time(RenewableProduction)))
RenewableProduction_month <- as.integer(format(time(RenewableProduction)))
#chat GPT prompt related to error for the type of date variable

# There has got to be a better way to do this
RenewableProduction_data <- data.frame(
  Year = RenewableProduction_year,
  Month = RenewableProduction_month,
  Production = as.numeric((RenewableProduction))
)

# Aggregate the series by year and month
RenewableProduction_matrix <- matrix(
  RenewableProduction_data$Production,
  nrow = 12, byrow = FALSE
)
```

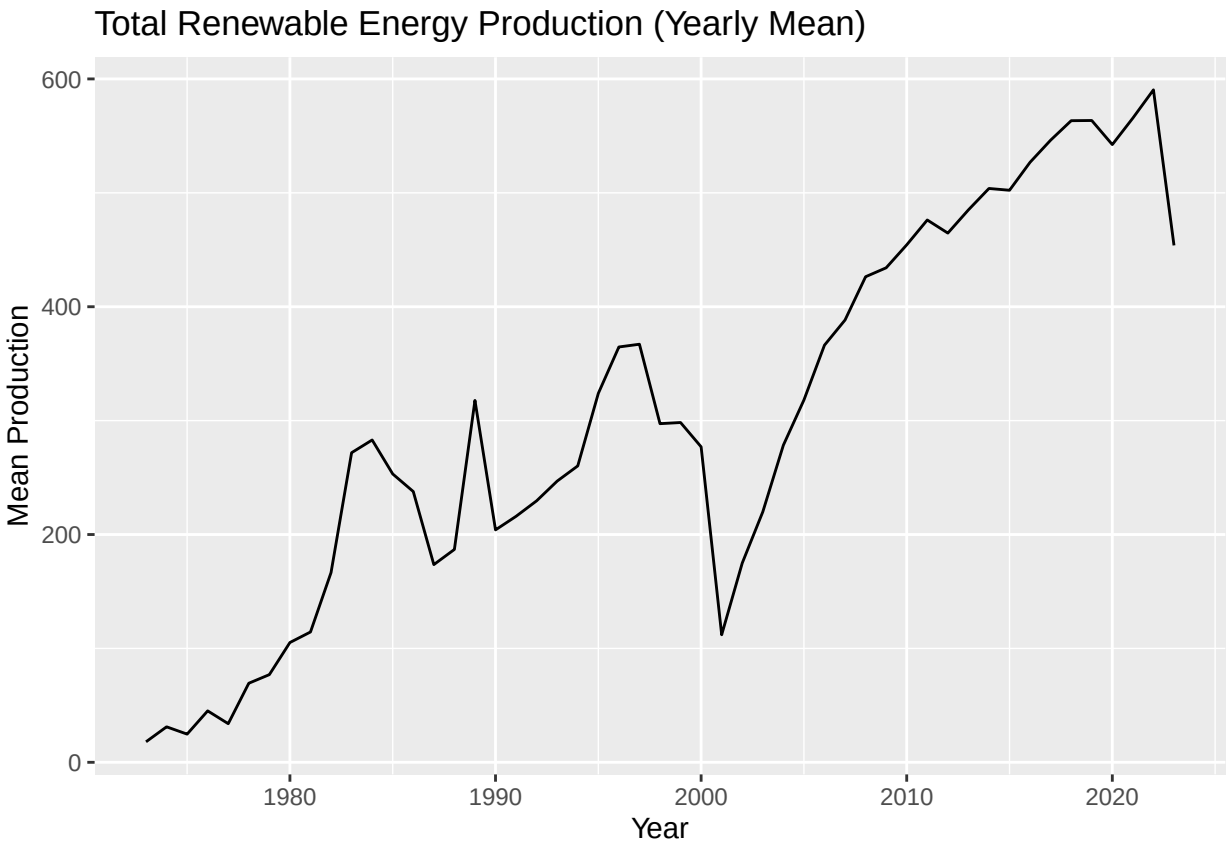
```
## Warning in matrix(RenewableProduction_data$Production, nrow = 12, byrow =
## FALSE): data length [609] is not a sub-multiple or multiple of the number of
## rows [12]
```

```r
# Take the column mean for each year
RenewableProduction_yearly_mean <- colMeans(RenewableProduction_matrix)

# Convert the yearly mean series into a time series object
RenewableProduction_time_series <- ts(
  RenewableProduction_yearly_mean,
  start = year1,
  frequency = 1
)

autoplot(RenewableProduction_time_series,
         main = "Total Renewable Energy Production (Yearly Mean)",
```

```
         xlab = "Year", ylab = "Mean Production")
```



Q7

Apply the Mann Kendal, Spearman correlation rank test and ADF. Are the results from the test in agreement with the test results for the monthly series, i.e., results for Q6?

```
print("Results for ADF test/n")
```

```
## [1] "Results for ADF test/n"
```

```
print(adf.test(RenewableProduction_time_series,alternative = "stationary"))
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  RenewableProduction_time_series
## Dickey-Fuller = -3.017, Lag order = 3, p-value = 0.1655
## alternative hypothesis: stationary
```

The ADF p-value is ~0.1665

```
SMKtest2 <- SeasonalMannKendall(RenewableProduction_time_series)
print("Results for Seasonal Mann Kendall /n")
```

```
## [1] "Results for Seasonal Mann Kendall /n"
```

```
print(summary(SMKtest2))
```

```
## Score =  1011 , Var(Score) = 15158.33
## denominator =  1275
## tau = 0.793, 2-sided pvalue =2.2204e-16
## NULL
```

The p-value is ~0

```
my_year <- RenewableProduction_data$Year
my_year <- unique(RenewableProduction_data$Year)
sp_rho <- cor(RenewableProduction_time_series, my_year, method = "spearman")
print("Results from Spearman Correlation:")
```

```
## [1] "Results from Spearman Correlation:"
```

```
print(sp_rho)
```

```
## [1] 0.9102262
```

```
sp_rho1=cor.test(RenewableProduction_time_series, my_year, method = "spearman")
print(sp_rho1)
```

```
##
##  Spearman's rank correlation rho
##
## data:  RenewableProduction_time_series and my_year
## S = 1984, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.9102262
```

The rho correlation value is 0.9102262 and the p-value is ~0

There still seems to be a positive trend on this data.