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

Assignment 4 - Due date 02/12/24

Aditi Jackson

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.\

```
#Load/install required package here
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

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

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

```
library(Kendall)
library(tseries)
library(readxl)
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(base)
library(cowplot)
```

```
##
## Attaching package: 'cowplot'
```

```
## The following object is masked from 'package:lubridate':
##
##     stamp
```

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".

```
# checking working directory
getwd()
```

```
## [1] "/home/guest/ENV797_APJ_S24_NEW/Assignments/RMD"
```

```
# had issues knitting due to using relative path; changed to absolute path and it worked
# loading data using read.csv
renewable_energy_full <-
  read.csv(
  "/home/guest/ENV797_APJ_S24_NEW/Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source
  header = TRUE, dec = ".", sep=",",stringsAsFactors = TRUE)

# renaming the "Month" column to "Date"
colnames(renewable_energy_full)[colnames(renewable_energy_full) == "Month"] <- "Date_MY"

# converting the "Date" column to a date object using lubridate
renewable_energy_full$Date_MY <- ym(renewable_energy_full$Date_MY)

# creating subset of data with Total Renewable Energy Production
## and Hydroelectric Power Consumption by date
renewable_energy_sub <- renewable_energy_full %>%
  select(
    Date_MY,
    Total.Renewable.Energy.Production)

# Verifying data
head(renewable_energy_sub)
```

```
##      Date_MY Total.Renewable.Energy.Production
## 1 1973-01-01                           219.839
## 2 1973-02-01                           197.330
## 3 1973-03-01                           218.686
## 4 1973-04-01                           209.330
## 5 1973-05-01                           215.982
## 6 1973-06-01                           208.249
```

```r
# transforming data into ts object
# start date is Jan 1, 1973
# frequency is 12 (monthly data)
renewable_energy_ts <- ts(renewable_energy_sub,start=c(1973,1),frequency=12)
```

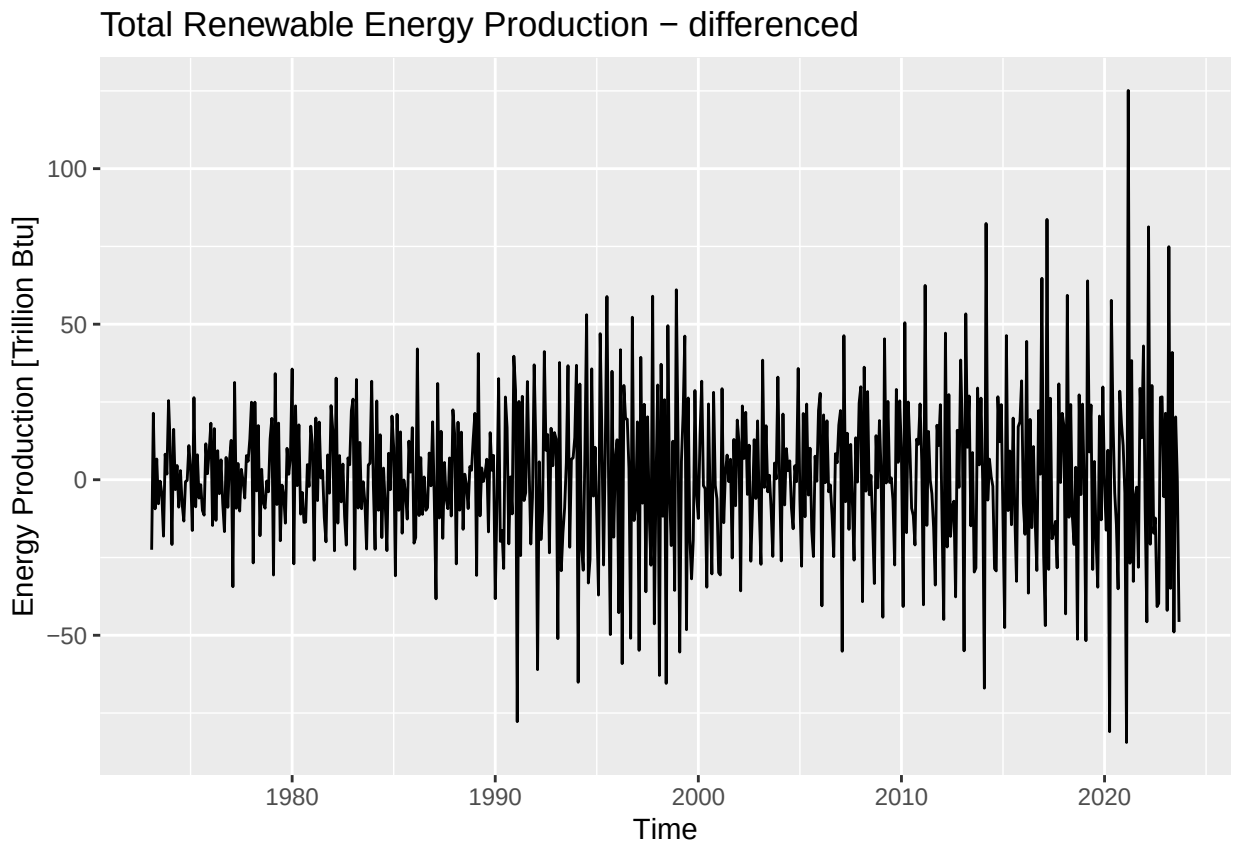
Stochastic Trend and Stationarity Tests

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?

```r
# differencing the data
Renew_diff <- diff(renewable_energy_ts[,2],lag=1,differences = 1)

# creating time series object
Renew_diff_ts=ts(Renew_diff, frequency=12,start=c(1973,1))

autoplot(Renew_diff)+
  ggtitle("Total Renewable Energy Production - differenced") +
  xlab("Time") +
  ylab("Energy Production [Trillion Btu]")
```



It seems like there is still some sort of trend remaining given similarity in wave patterns at yearly intervals - perhaps some sort of seasonality.

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.

```
# creating vector "t" to represent number of observations
num_obs <- nrow(renewable_energy_sub)
t <- 1:num_obs

# fitting linear regression for Renewable Energy Consumption time series
Renewables_linReg <-
  lm(Total.Renewable.Energy.Production ~ t, data = renewable_energy_sub)
summary(Renewables_linReg)
```

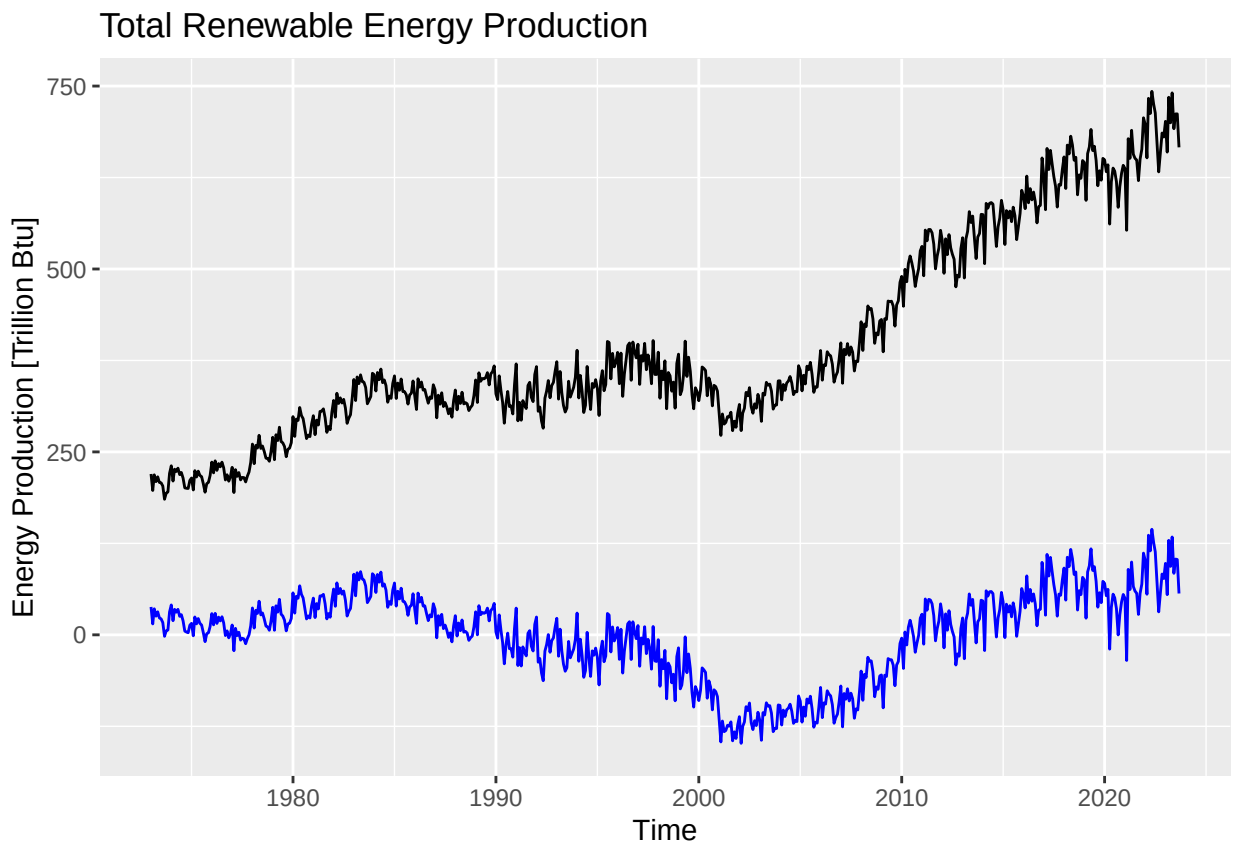
```
##
## Call:
## lm(formula = Total.Renewable.Energy.Production ~ t, data = renewable_energy_sub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -148.27  -35.63   11.58   41.51  144.27
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 180.98940    4.90151   36.92   <2e-16 ***
## t             0.70404    0.01392   50.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.41 on 607 degrees of freedom
## Multiple R-squared:  0.8081, Adjusted R-squared:  0.8078
## F-statistic:  2557 on 1 and 607 DF,  p-value: < 2.2e-16
```

```
# saving coefficients
Renew_beta0 <- Renewables_linReg$coefficients[1]
Renew_beta1 <- Renewables_linReg$coefficients[2]

# de-trending renewable energy production time series
Renew_detrend <- renewable_energy_sub[,2] - (Renew_beta0+Renew_beta1*t)

# creating time series object
Renew_detrend_ts=ts(Renew_detrend, frequency=12,start=c(1973,1))

# creating data frame in order to plot
df_renew_detrend <-
  data_frame("date"=renewable_energy_sub$Date,
             "observed"=renewable_energy_sub[,2],
             "detrend"=Renew_detrend)
```

```
## Warning: `data_frame()` was deprecated in tibble 1.1.0.
## i Please use `tibble()` instead.
## This warning is displayed once every 8 hours.
```

```
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
# plotting detrended series
ggplot(df_renew_detrend,aes(x=date))+
  geom_line(aes(y=observed),color="black")+
  geom_line(aes(y=detrend),color="blue")+
  labs(x="Time",
       y="Energy Production [Trillion Btu]",
       title="Total Renewable Energy Production")
```

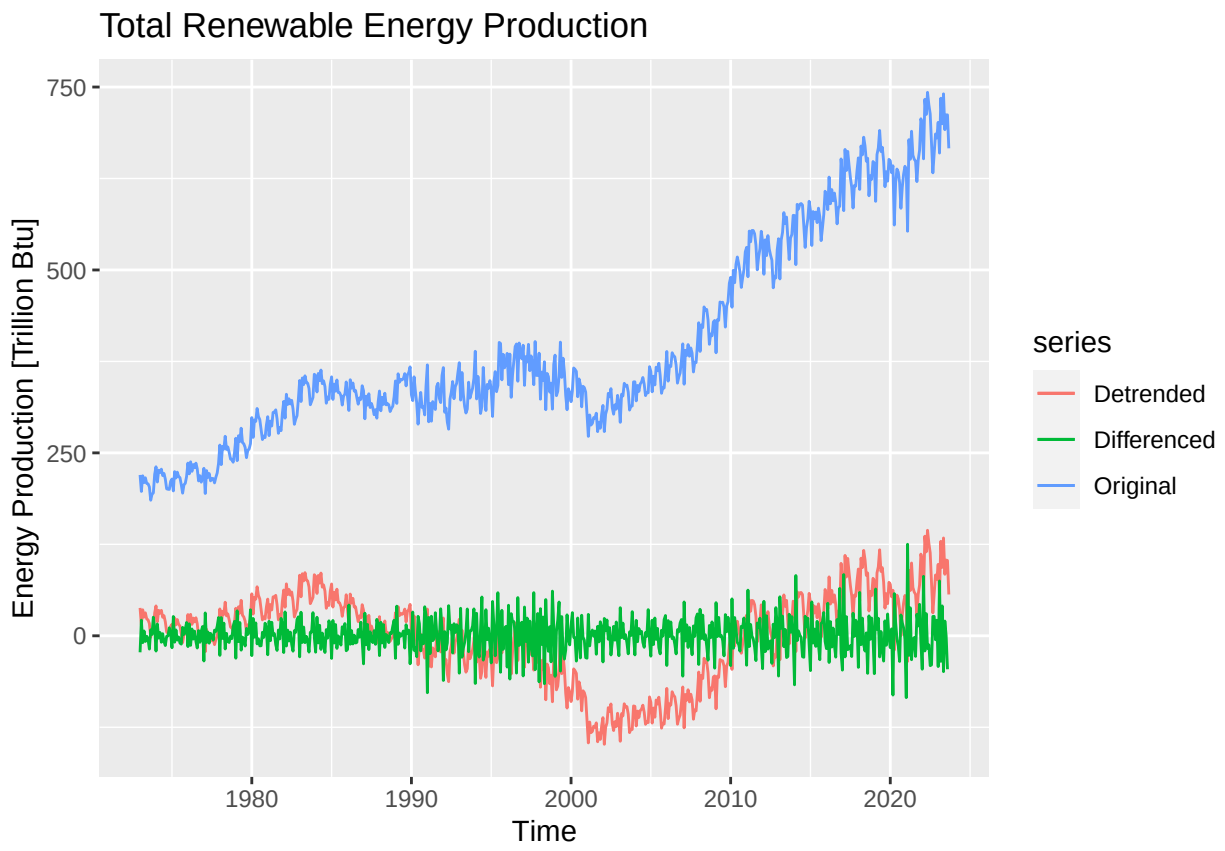


Q3

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.

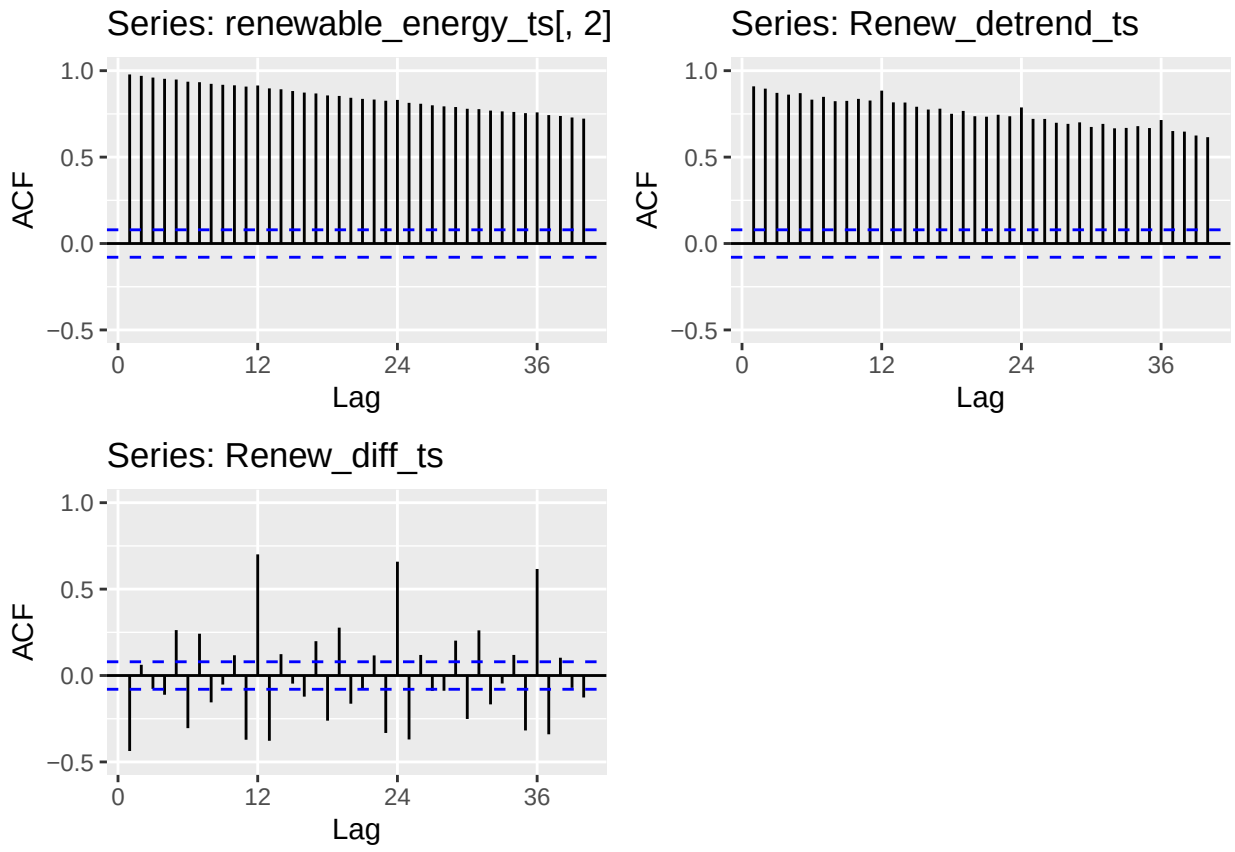
```
# plotting original, detrended, and differenced series
autoplot(renewable_energy_ts[,2],series="Original")+
  autolayer(Renew_detrend_ts,series="Detrended")+
  autolayer(Renew_diff_ts,series="Differenced")+
  ylab("Energy Production [Trillion Btu]")+
  ggtitle("Total Renewable Energy Production")
```



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?

```
# ACF of original series
ACF_original <-
  autoplot(Acf(renewable_energy_ts[,2],lag.max = 40,plot=FALSE))+ylim(-0.5, 1)

# ACF of detrended series
ACF_detredned <-
  autoplot(Acf(Renew_detrend_ts,lag.max=40,plot=FALSE))+ylim(-0.5, 1)

# ACF of differenced series
ACF_differenced <-
  autoplot(Acf(Renew_diff_ts, lag.max=40,plot=FALSE))+ylim(-0.5, 1)

# plotting all three on the same grid for easier comparison
plot_grid(ACF_original,ACF_detredned,ACF_differenced)
```



In this case it appears that differeing was more effective at removing the trend. This is because the magnitudes of the ACF values are decreased significantly compared to the original and linear regression series, and their seems to be less time dependence overall. This makes sense given that we are looking at a stochastic trend, which is not well approximated by a linear model.

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 Q2? 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.

```
# Seasonal Mann-Kendall Test
SMKtest <- SeasonalMannKendall(renewable_energy_ts[,2])
print("Results for Seasonal Mann Kendall /n")
```

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

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

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

```
# ADF Test
## null hypothesis is that data has a unit root (aka is stochastic)
print("Results for ADF test/n")
```

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

```
print(adf.test(renewable_energy_ts[,2],alternative = "stationary"))
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  renewable_energy_ts[, 2]
## Dickey-Fuller = -1.24, Lag order = 8, p-value = 0.9
## alternative hypothesis: stationary
```

The Seasonal Mann Kendal Test has a test statistic value (tau) of 0.783, which suggests a strong positive trend. The p-value is < 2.22e-16, which indicates that there is strong evidence against the null hypothesis that there is no trend. The concludion of this test is that is it unlikely no trend exists.

The ADF Test has a test statistic of -1.24 and a p-value of 0.9. Since the p-value is greater than alpha = 0.05, we fail to reject the null hypothesis that the data has a unit root. This suggests the presence of a stochastic trend.

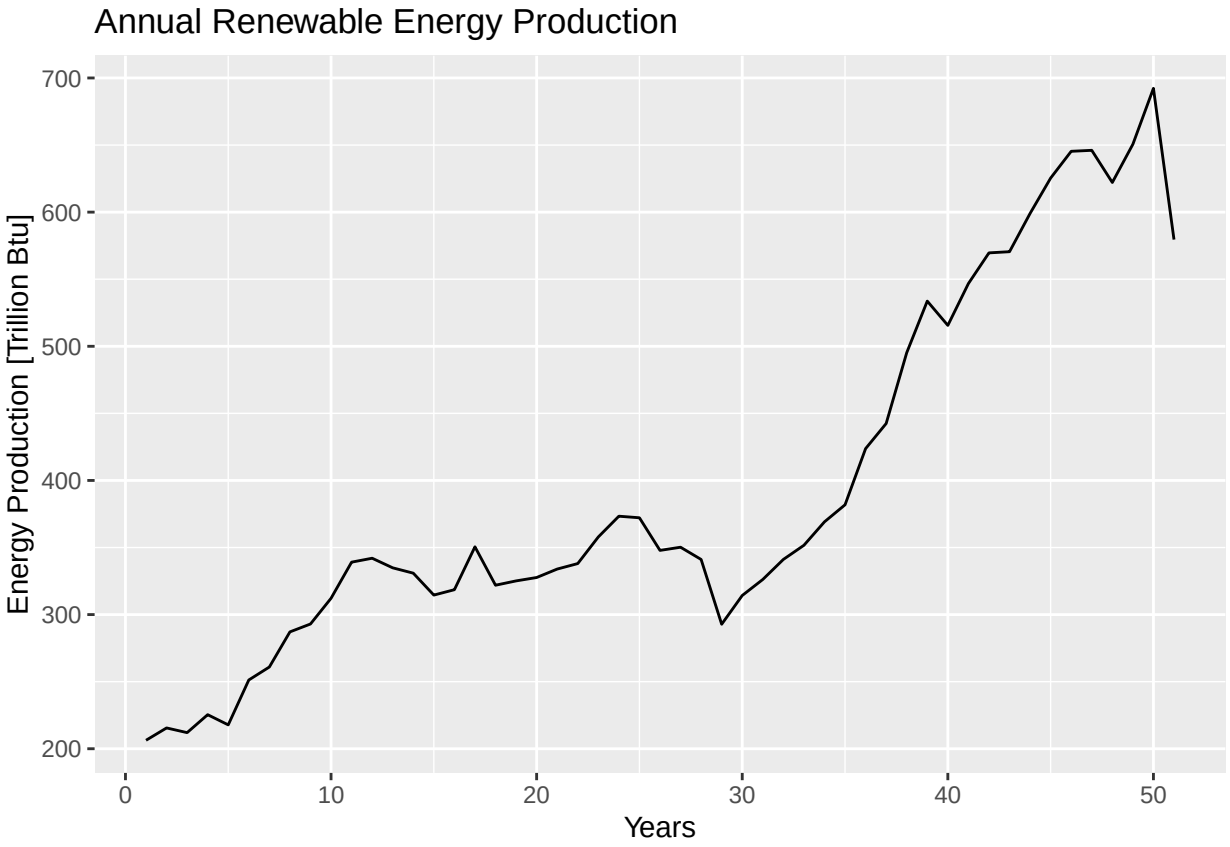
Q6

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().

```
# grouping data into yearly steps using matrix method
renewables_matrix <- matrix(renewable_energy_ts[,2],byrow=FALSE,nrow=12)
```

```
## Warning in matrix(renewable_energy_ts[, 2], byrow = FALSE, nrow = 12): data
## length [609] is not a sub-multiple or multiple of the number of rows [12]
```

```
renewables_by_year <- colMeans(renewables_matrix)

# defining years
Date_MY <- renewable_energy_sub$Date_MY
Year_MY <- c(year(first(Date_MY)):year(last(Date_MY)))

# grouping data into dataframe
renewables_by_year_NEW <- data.frame(Year_MY, renewables_by_year)

# converting to time series object
renewables_yearly_ts <- ts(renewables_by_year_NEW)

# plotting ts
autoplot(renewables_yearly_ts[,2])+
  xlab("Years")+
  ylab("Energy Production [Trillion Btu]")+
  ggtitle("Annual Renewable Energy 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?

```
# Seasonal Mann-Kendall Test
SMKtest <- SeasonalMannKendall(renewables_yearly_ts[,2])
print("Results for Seasonal Mann Kendall /n")
```

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

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

```
## Score =  1019 , Var(Score) = 15158.33
## denominator =  1275
## tau = 0.799, 2-sided pvalue =2.2204e-16
## NULL
```

```
# Spearman Correlation Test
print("Results from Spearman Correlation")
```

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

```
sp_rho=cor(renewables_yearly_ts[,2],Year_MY,method="spearman")
print(sp_rho)
```

```
## [1] 0.9136652
```

```
# ADF Test
## null hypothesis is that data has a unit root (aka is stochastic)
```

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

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

```r
print(adf.test(renewables_yearly_ts[,2],alternative = "stationary"))
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  renewables_yearly_ts[, 2]
## Dickey-Fuller = -2.0953, Lag order = 3, p-value = 0.5361
## alternative hypothesis: stationary
```

Yes, the results in Q7 align with the results from Q5 and Q6. It appears that there is some sort of non-stationary trend present: > With the yearly data, the Seasonal Mann-Kendal Test has a tau of 0.799 and a p-value of -2.2204e-16, which provide strong evidence against the null hypothesis that there is no trend. This therefore suggests that there may be some sort of trend present in the yearly production data. > The Spearman correlation coefficient is ~0.914, indicating a strong positive relationship between time and annual renewable energy production. > The ADF test has a test stat of -2.0943 and a p-value of 0.5361. Since the p-value is greater than a significance level of alpha = 0.05, we fail to reject the null hypothesis that the yearly series is non-stationary. The conclusion would be that there is likely a stochastic trend present.