ENV 790.30 - Time Series Analysis for Energy Data | Spring 2022

Assignment 4 - Due date 02/17/22

Student Name

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 project open the first thing you will do is change "Student Name" on line 3 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. Rename the pdf file such that it includes your first and last name (e.g., "LuanaLima_TSA_A04_Sp21.Rmd"). 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(forecast)
library(tseries)
library(Kendall)
library(dplyr)
library(tidyverse)
library(readxl)
library(ggfortify)
library(ggplot2)
library(zoo)
```

Questions

Consider the same data you used for A3 from the spreadsheet "Table_10.1_Renewable_Energy_Production_and_Consumpti
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 xlsx package
energy <- read_excel("~/ENVIRON 790/ENV790_TimeSeriesAnalysis_Sp2022/Data/Table_10.1_Renewable_Energy_P
                     skip = 10)

energy <- energy %>%
  slice(2:nrow(energy)) %>% # delete extraneous first row
```

```
  select(`Total Renewable Energy Production`) %>%
  mutate(`Total Renewable Energy Production` = as.numeric(`Total Renewable Energy Production`))
energy_ts <- ts(energy, start = c(1973, 1), frequency = 12)
```

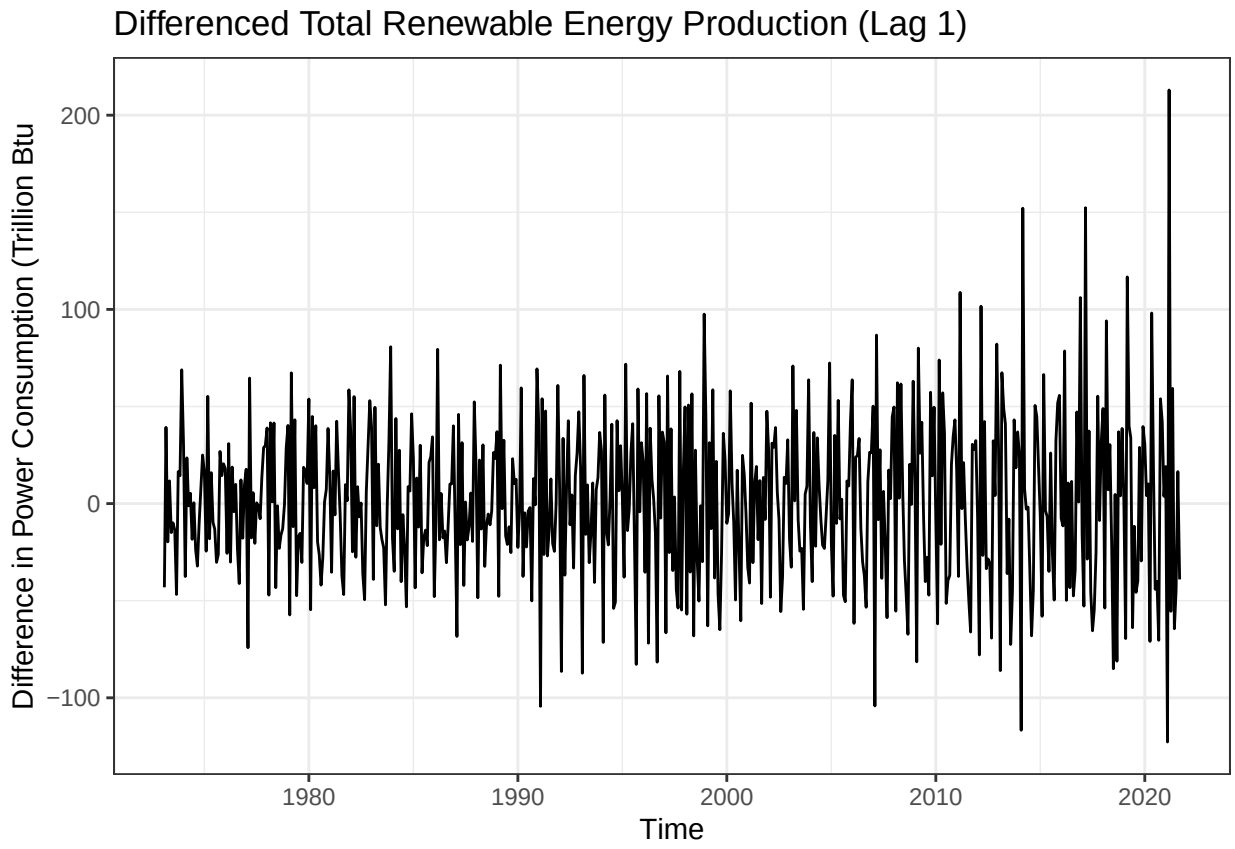
Stochastic Trend and Stationarity Tests

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?

```
diff_series <- diff(energy$`Total Renewable Energy Production`, lag = 1, differences = 1)
diff_ts <- ts(diff_series, start = c(1973, 2), frequency = 12)
autoplot(diff_ts) +
  labs(title = "Differenced Total Renewable Energy Production (Lag 1)",
       x = "Time", y = "Difference in Power Consumption (Trillion Btu") +
  theme_bw()
```



Looking at the plot of the differenced series, we can see that there no longer seems to be a particular trend (increasing or decreasing) in the data.

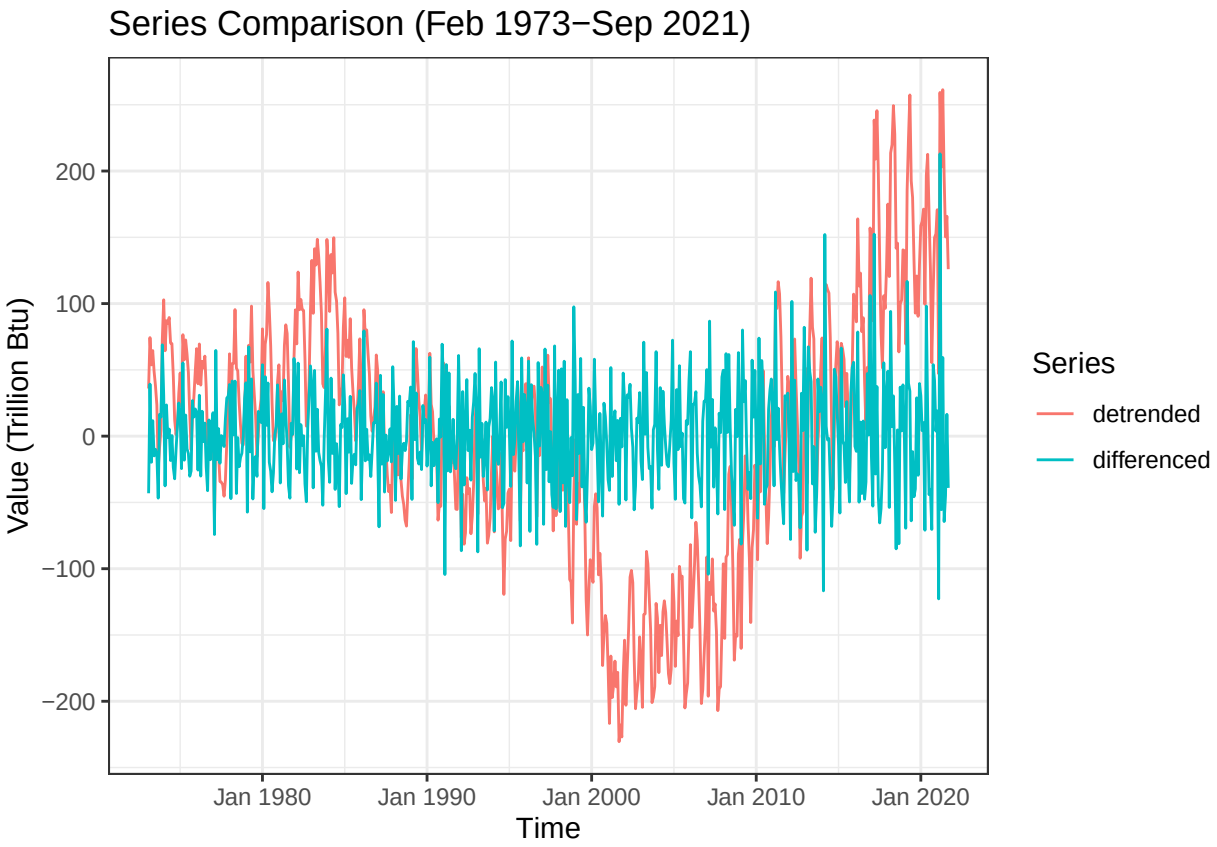
Q2

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 A3 using linear regression. (Hint: Just copy and paste part of your code for A3)

Copy and paste part of your code for A3 where you compute regression for Total Energy Production and the detrended Total Energy Production

```
t <- c(1:nrow(energy_ts))
lm_mod <- lm(energy_ts ~ t)
beta0 <- as.numeric(lm_mod$coefficients[1])
beta1 <- as.numeric(lm_mod$coefficients[2])
detrend_series <- energy_ts - (beta0 + beta1 * t)

tibble(month = as.yearmon(time(diff_ts)),
       detrended = detrend_series[2:585],
       differenced = diff_ts) %>%
  pivot_longer(detrended:differenced, names_to = "series") %>%
  ggplot(aes(x = month, y = value, color = series)) +
    geom_line() +
    labs(title = "Series Comparison (Feb 1973-Sep 2021)",
         x = "Time", y = "Value (Trillion Btu)", color = "Series") +
    theme_bw()
```



Comparing the two series, it seems like the differenced series was better at eliminating the trend. Although we were able to reduce the slope to zero using linear regression, it is still evident from the detrended series

that there are noticeable fluctuations (dips and peaks), whereas the fluctuations for the differenced series remain relatively constant.

Q3

Create a data frame with 4 columns: month, original series, detrended by Regression Series and differenced series. Make sure you properly name all columns. Also note that the differenced series will have only 584 rows because you loose the first observation when differencing. Therefore, you need to remove the first observations for the original series and the detrended by regression series to build the new data frame.

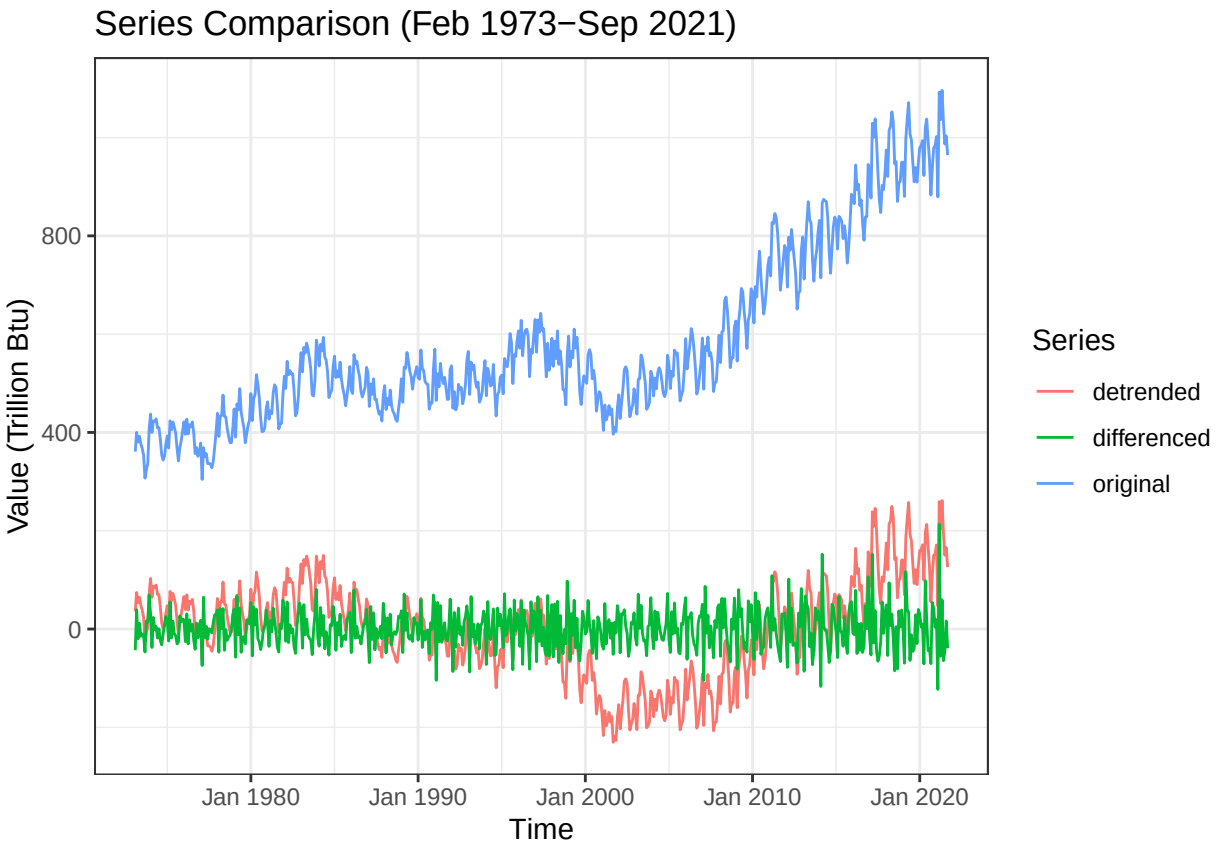
```
#Data frame - remember to note include January 1973
all_series <- tibble(month = as.yearmon(time(diff_ts)),
                     original = energy_ts[2:585],
                     detrended = detrend_series[2:585],
                     differenced = diff_ts)
head(all_series)
```

```
## # A tibble: 6 x 4
##   month     original detrended differenced
##   <yearmon>    <dbl>     <dbl>       <dbl>
## 1 Feb 1973      361.      36.0      -43.1
## 2 Mar 1973      400.      74.3       39.3
## 3 Apr 1973      380.      53.8      -19.7
## 4 May 1973      392.      64.6       11.7
## 5 Jun 1973      377.      48.8      -14.9
## 6 Jul 1973      367.      38.0       -9.91
```

Q4

Using ggplot() create a line plot that shows the three series together. Make sure you add a legend to the plot.

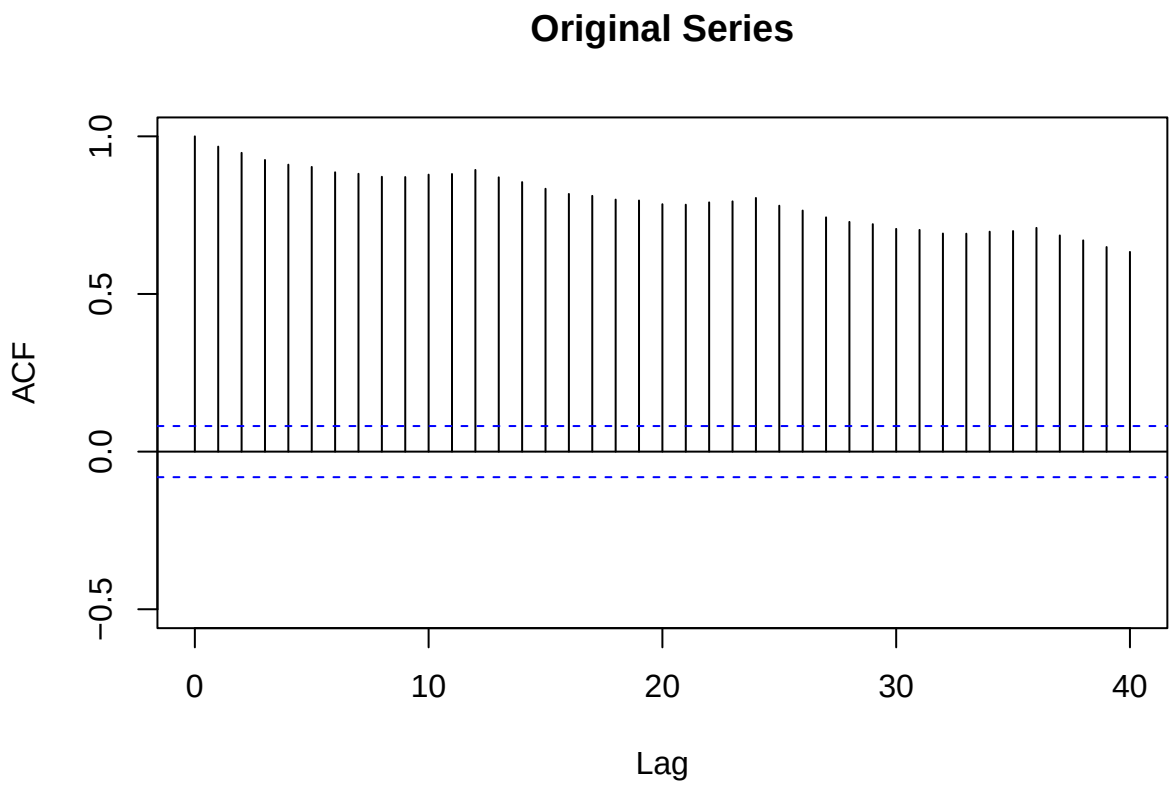
```
#Use ggplot
all_series %>%
  pivot_longer(original:differenced, names_to = "series") %>%
  ggplot(aes(x = month, y = value, color = series)) +
    geom_line() +
    labs(title = "Series Comparison (Feb 1973-Sep 2021)",
         x = "Time", y = "Value (Trillion Btu)", color = "Series") +
    theme_bw()
```



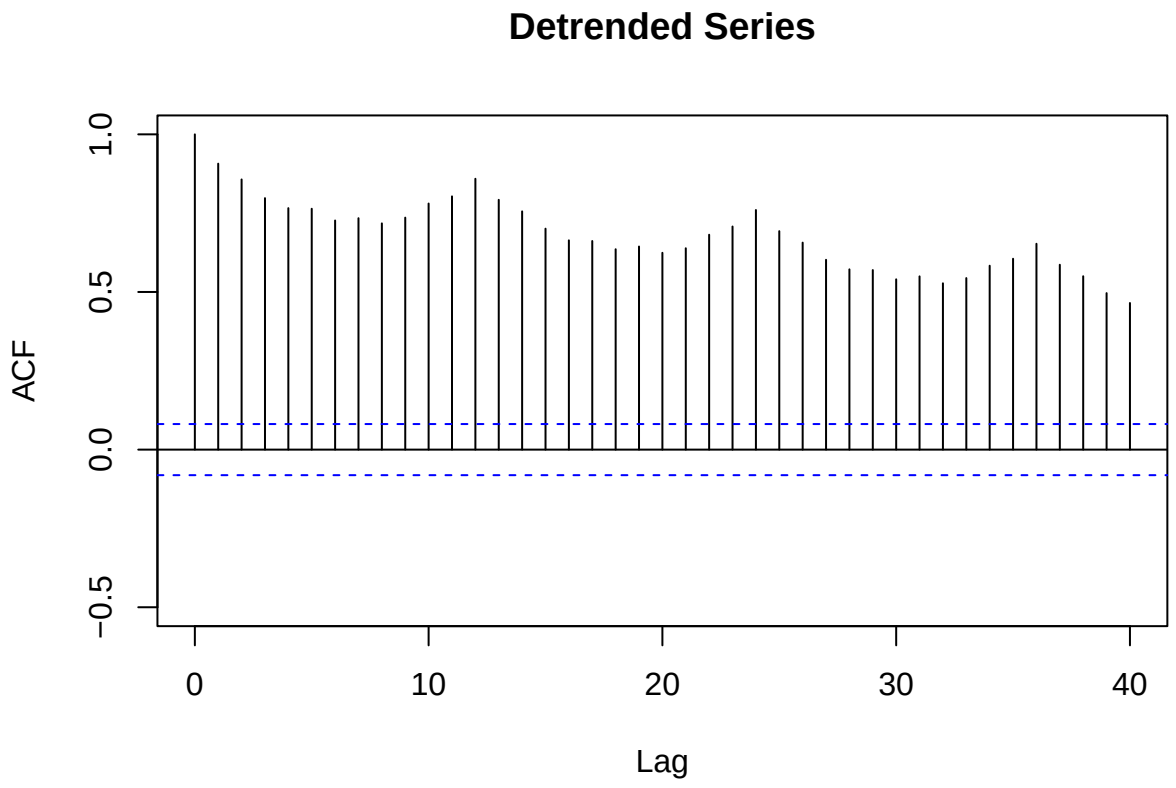
Q5

Plot the ACF for the three series and compare the plots. Add the argument `ylim=c(-0.5,1)` to the Acf() function 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?

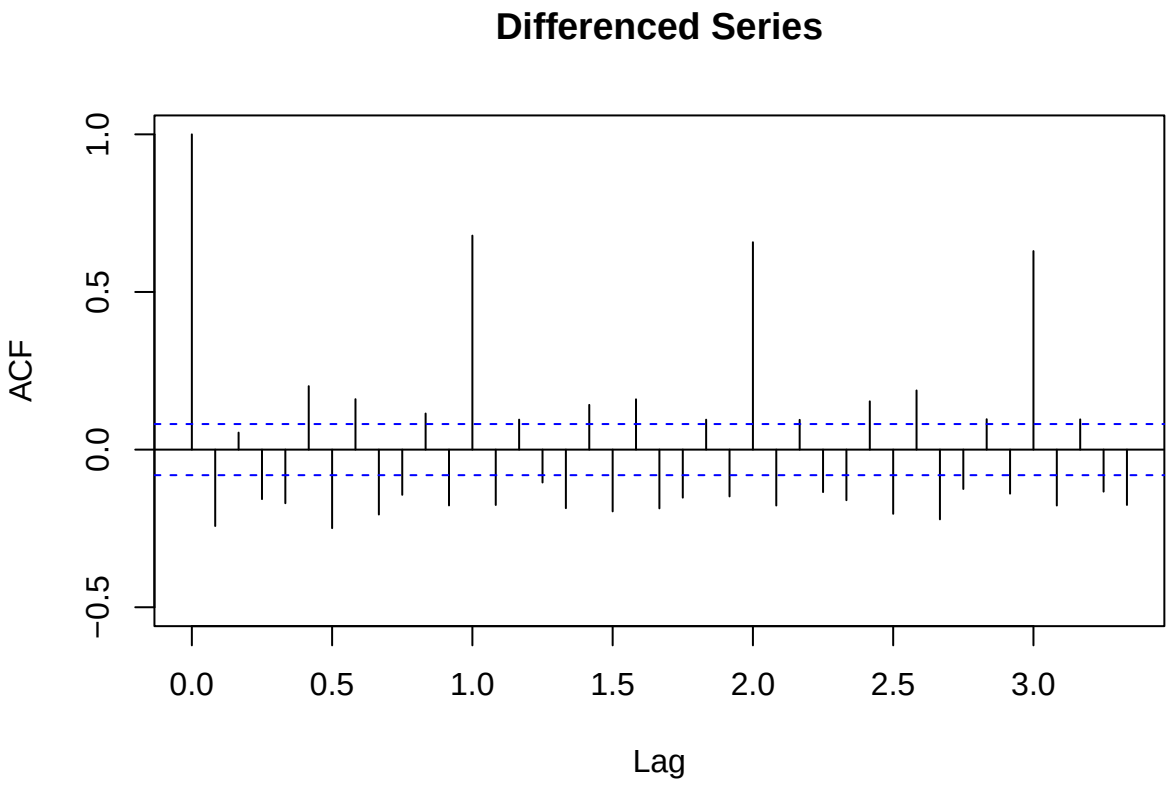
```
#Compare ACFs
acf(all_series$original, lag.max = 40, main = "Original Series", ylim=c(-0.5,1))
```




```
acf(all_series$detrended, lag.max = 40, main = "Detrended Series", ylim=c(-0.5,1))
```




```
acf(all_series$differenced, lag.max = 40, main = "Differenced Series", ylim=c(-0.5,1))
```



Differencing was more effective at eliminating the trend, as we can see from the ACF plots that the plot for differencing shows an immediate and significant drop in the ACF values after the first lag. On the other hand, both plots for the original and the detrended series show much more gradual reductions in the ACF values as lag increases; although we do see more of a greater decrease in values for the detrended series, the difference is rather minimal when compared to the original series and still much more of a gradual decrease when compared to the differenced series.

Q6

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. Whats 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.

```
SMKtest <- SeasonalMannKendall(energy_ts)
print("Results for Seasonal Mann Kendall")
```

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

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

```
## Score =  9984 , Var(Score) = 159104
## denominator =  13968
```

```
## tau = 0.715, 2-sided pvalue =< 2.22e-16
## NULL
```

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

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

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

```
##
##  Augmented Dickey-Fuller Test
##
## data:  energy_ts
## Dickey-Fuller = -1.4383, Lag order = 8, p-value = 0.8161
## alternative hypothesis: stationary
```

After performing the Seasonal Mann-Kendall test, the calculated p-value is $2.22 \times 10^{-16} < 0.05$, therefore we reject the null hypothesis and conclude that there is a deterministic trend. After performing the ADF test, the calculated p-value is $0.8161 > 0.05$, therefore we fail to reject the null hypothesis and conclude that the data has a unit root and therefore the series has a stochastic trend.

This matches what we observed in Q2, where we saw that differencing the series seemed to do a much better job at removing the trend in the series as compared to detrending it using linear regression. Since the ADF test shows that the series has a stochastic trend, a different procedure other than linear regression is necessary to remove the trend.

Q7

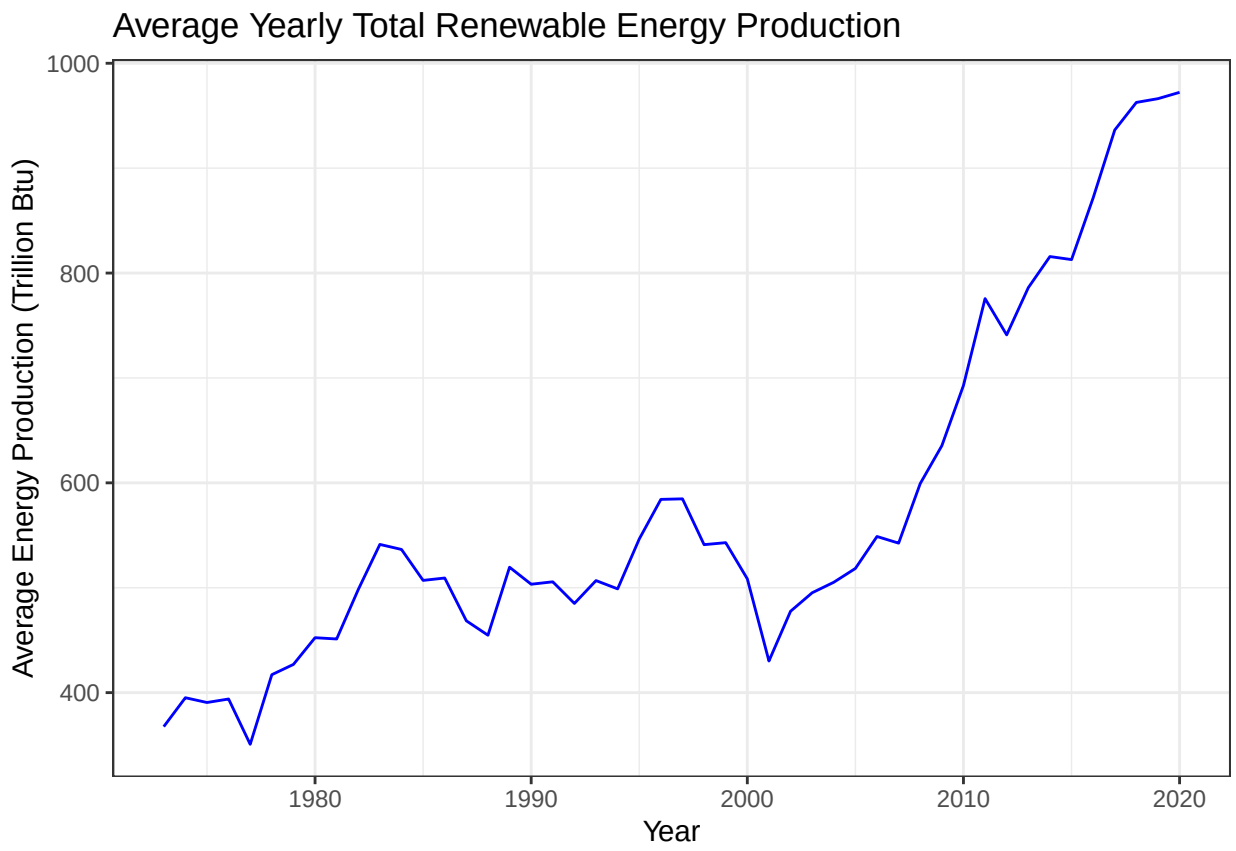
Aggregate the original "Total Renewable Energy Production" series by year. You can use the same procedure we used in class. Store series in a matrix where rows represent months and columns represent years. And then take the columns mean using function colMeans(). Recall the goal is the remove the seasonal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

```
# remove 2021 since the data for that year is incomplete
energy_ts <- ts(energy[1:576,], start = c(1973, 1), frequency = 12)

energy_matrix <- matrix(energy_ts, byrow=FALSE, nrow=12)
energy_yearly <- colMeans(energy_matrix)
year <- as.numeric(unique(format(as.yearmon(time(energy_ts)), "%Y")))

energy_new_yearly <- data.frame(year, energy_yearly)

ggplot(energy_new_yearly, aes(x=year, y=energy_yearly)) +
  geom_line(color="blue") +
  labs(title = "Average Yearly Total Renewable Energy Production",
       x = "Year", y = "Average Energy Production (Trillion Btu)") +
  theme_bw()
```



Q8

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

```
print("Results of Mann Kendall on average yearly series")
```

```
## [1] "Results of Mann Kendall on average yearly series"
```

```
print(summary(MannKendall(energy_yearly)))
```

```
## Score =  816 , Var(Score) = 12658.67
## denominator =  1128
## tau = 0.723, 2-sided pvalue =< 2.22e-16
## NULL
```

```
sp_rho <- cor.test(energy_yearly, year, method="spearman")
print(sp_rho)
```

```
##
##  Spearman's rank correlation rho
##
## data:  energy_yearly and year
```

```
## S = 2548, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8617021
```

```
print("Results for ADF test on yearly data")
```

```
## [1] "Results for ADF test on yearly data"
```

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

```
##
##  Augmented Dickey-Fuller Test
##
## data:  energy_yearly
## Dickey-Fuller = -1.0426, Lag order = 3, p-value = 0.9219
## alternative hypothesis: stationary
```

Yes, we can see that the results from these tests are in agreement with the test results of the non-aggregated series as seen in Q6, as both reject the null hypothesis for the Mann Kendall test and fail to reject the null for the ADF test.