ENV 797 - Time Series Analysis for Energy and Environment Applications | Spring 2026

Assignment 4 - Due date 02/10/26

Joshua Salzberg

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_Sp26.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
librarylist <- c("openxlsx", "readxl",'ggplot2','forecast','tseries','Kendall','here', 'tidyverse')
lapply(librarylist, require, character.only=TRUE)
```

```
## [[1]]
## [1] TRUE
##
## [[2]]
## [1] TRUE
##
## [[3]]
## [1] TRUE
##
## [[4]]
## [1] TRUE
##
## [[5]]
## [1] TRUE
##
## [[6]]
## [1] TRUE
##
## [[7]]
```

```
## [1] TRUE
##
## [[8]]
## [1] TRUE
```

```
here()
```

```
## [1] "C:/Users/jhsal/OneDrive - Duke University/797/TSA2026/TSA_Sp26"
```

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 December 2025 Monthly Energy Review. **For this assignment you will work only with the column "Total Renewable Energy Production"**.

```
#Importing data set - you may copy your code from A3
energydata <- read.csv(here("Data", "Processed","jhs_clean_energy_data.csv"))
# Selecting energy data
A4data <- energydata %>%
  select(TotalRenewables = Total.Renewable.Energy.Production)

A4ts <- ts(A4data, start = c(1973, 1), frequency = 12)
```

Stochastic Trend and Stationarity Tests

For this part you will work only with the column Total Renewable Energy Production.

Q1

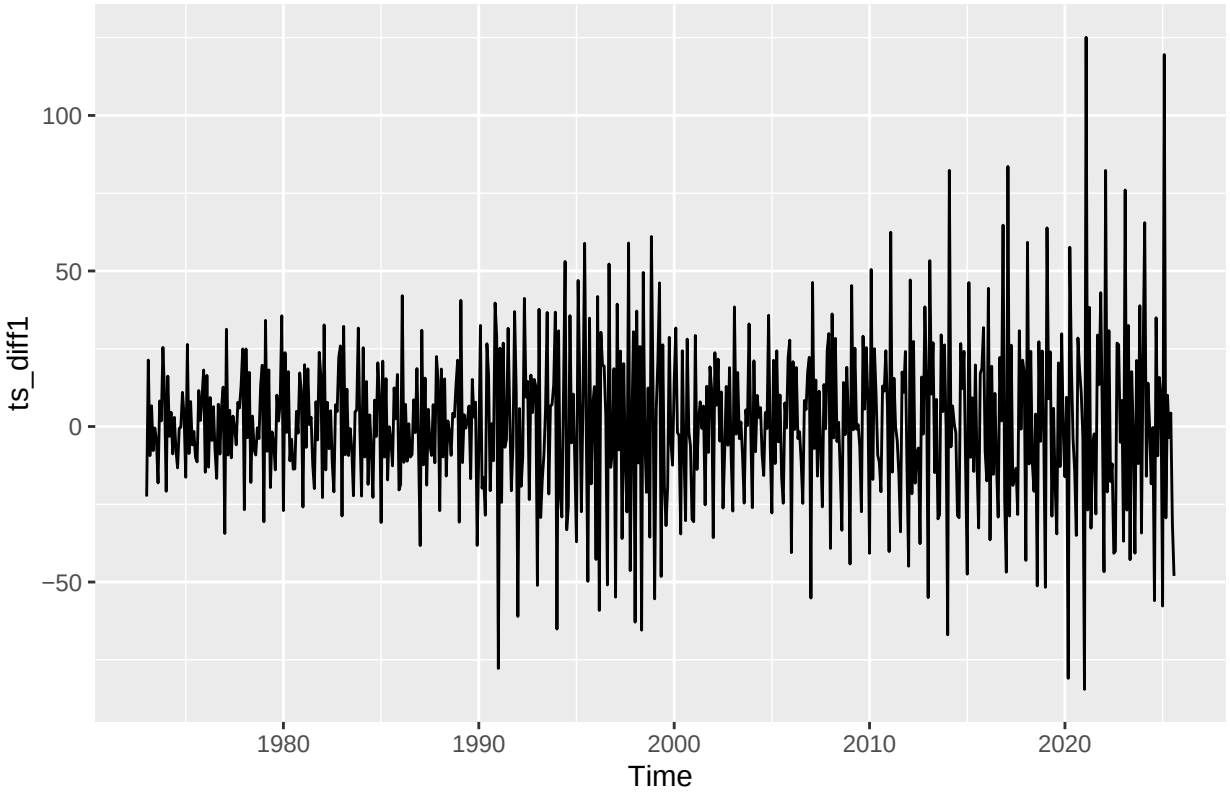
Difference the "Total Renewable Energy Production" series using function diff(). Function diff() is from package base and take three main arguments: * *x* vector containing values to be differenced; * *lag* integer indicating with lag to use; * *differences* integer indicating how many times series should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

```
# difference the data
diff1 <- diff(A4data$TotalRenewables, lag = 1, differences = 1)

# make it time series data
ts_diff1 <- ts(data = diff1, start = c(1973, 1), frequency = 12)

#plot the data
autoplot(ts_diff1)
```



answer: the series appears to have no trend.

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 original series. This should be the code for Q3 and Q4. make sure you use assign same name for the time series object that you had in A3, otherwise the code will not work.

```
#Vector for linear regression
t <- c(1:nrow(A4data))

# renewables regression and coefficients
LM1 <- lm(A4data[,1]~t )
summary(LM1)
```

```
##
## Call:
## lm(formula = A4data[, 1] ~ t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -154.81  -39.55   12.52   41.49  171.15
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 171.44868    5.11085   33.55   <2e-16 ***
## t             0.74999    0.01397   53.69   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64.22 on 631 degrees of freedom
## Multiple R-squared:  0.8204, Adjusted R-squared:  0.8201
## F-statistic:  2883 on 1 and 631 DF,  p-value: < 2.2e-16
```

```
LM1Beta0 <- as.numeric(LM1$coefficients[1]) # Intercept
LM1Beta1 <- as.numeric(LM1$coefficients[2]) # Slope
# Detrending Renewable Data
detrend_renewables <- A4data[,1]-(LM1Beta0+LM1Beta1*t)
class(detrend_renewables)
```

```
## [1] "numeric"
```

```
ts_detrend_renewables <- ts(detrend_renewables, frequency=12, start=c(1973,1))
```

Q3

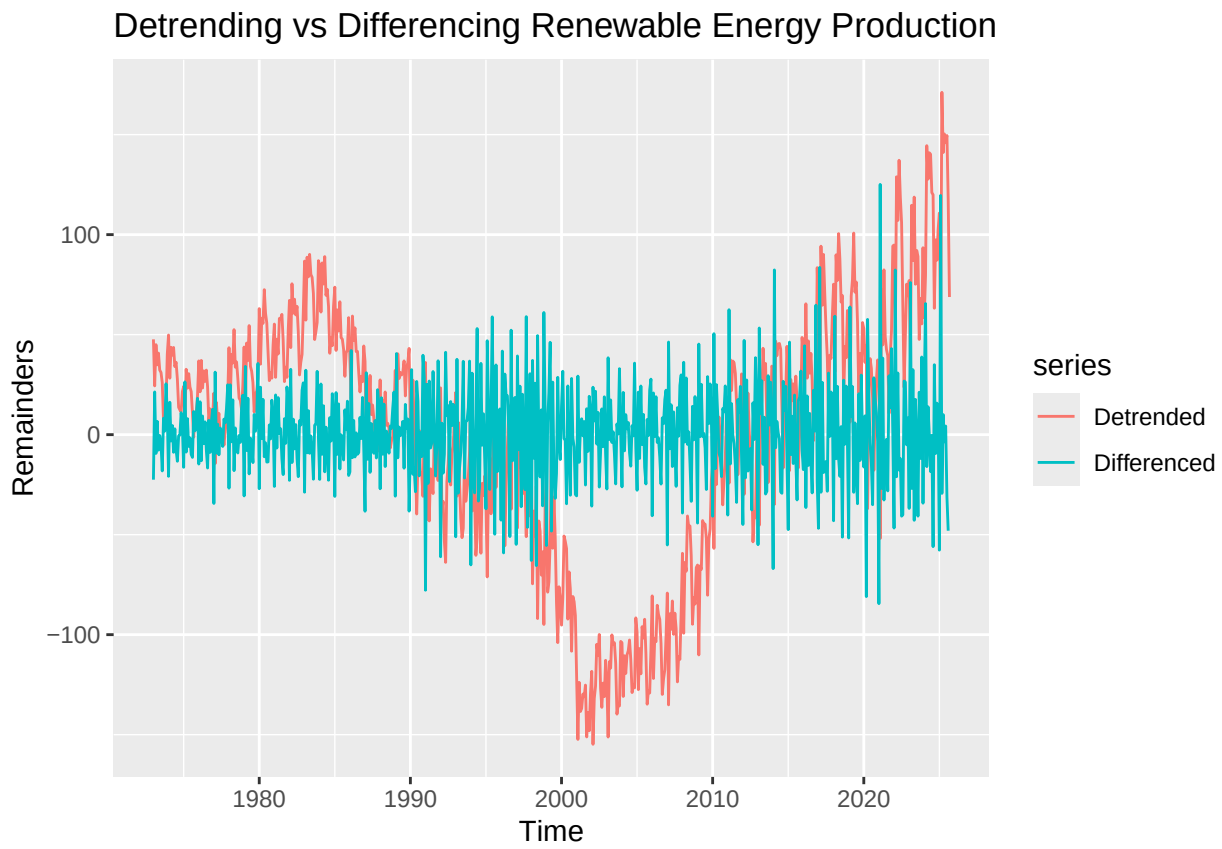
Now let's compare the differenced series with the detrended series you calculated on A3. In other words, for the "Total Renewable Energy Production" compare the differenced series from Q1 with the series you detrended in Q2 using linear regression.

Using autoplot() + autolayer() create a plot that shows the three series together (i.e. "Original", "Differenced", "Detrended lm()"). 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 on how to use autoplot() and autolayer().

What can you tell from this plot? Which method seems to have been more efficient in removing the trend?

```
# autoplot + autolayer

autoplot(ts_detrend_renewables, series = "Detrended")+
  autolayer(ts_diff1, series = "Differenced")+
  labs(title = "Detrending vs Differencing Renewable Energy Production",
       x = "Time",
       y = "Remainders")
```

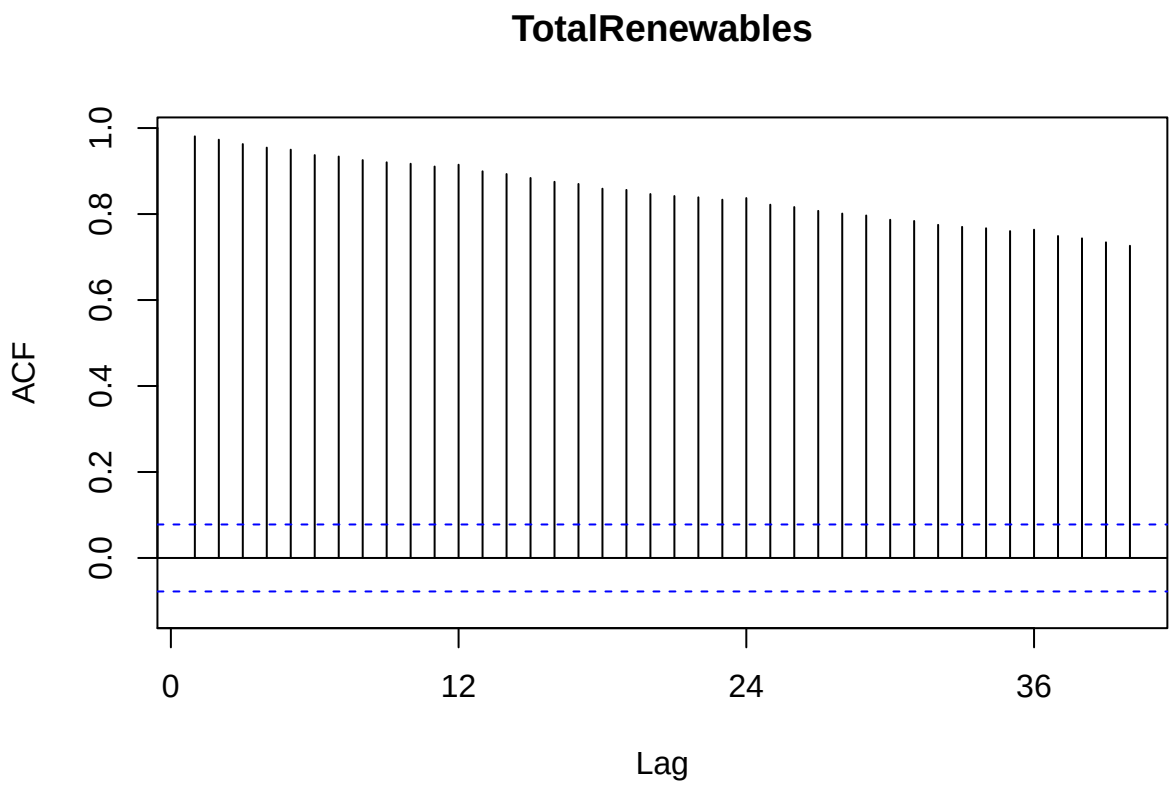


Answer: The Detrended series retained significant peaks and valleys, signifying significant trend remains in the series; it lacks the stationarity we would hope for resulting from the process. On the other hand, the Differenced series appears completely stationary, with very strong reversion back to the y=0 line, indicating the trend has been substantially removed.

Q4

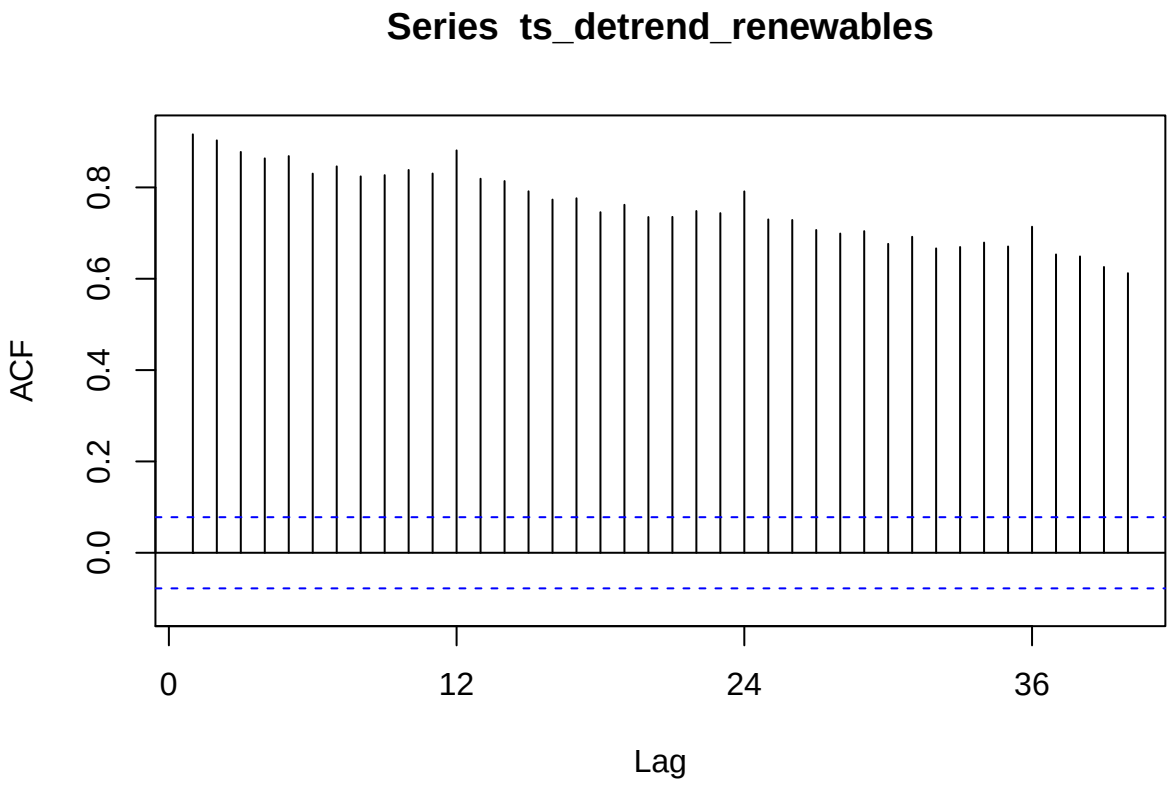
Plot the ACF for the three series and compare the plots. Add the argument `ylim=c(-0.5,1)` to the autoplot() or Acf() function - whichever you are using to generate the plots - to make sure all three y axis have the same limits. Looking at the ACF which method do you think was more efficient in eliminating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

```
# 3 ACF plots
Renewables1 <- autoplot(Acf(A4ts, lag.max = 40), ylim =c(-0.5,1))
```



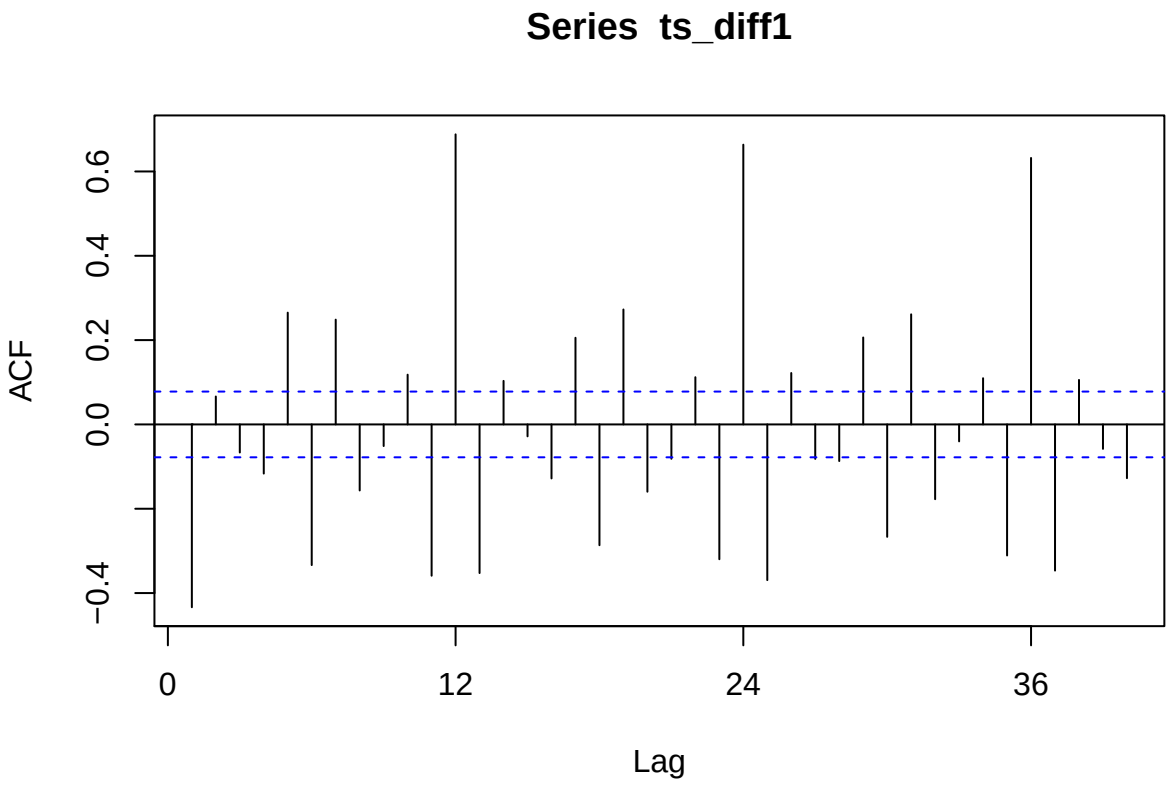

```
## Warning in ggplot2::geom_segment(lineend = "butt", ...): Ignoring unknown
## parameters: `ylim`
```

```
Renewables2 <- autoplot(Acf(ts_detrend_renewables, lag.max = 40), ylim =c(-0.5,1))
```



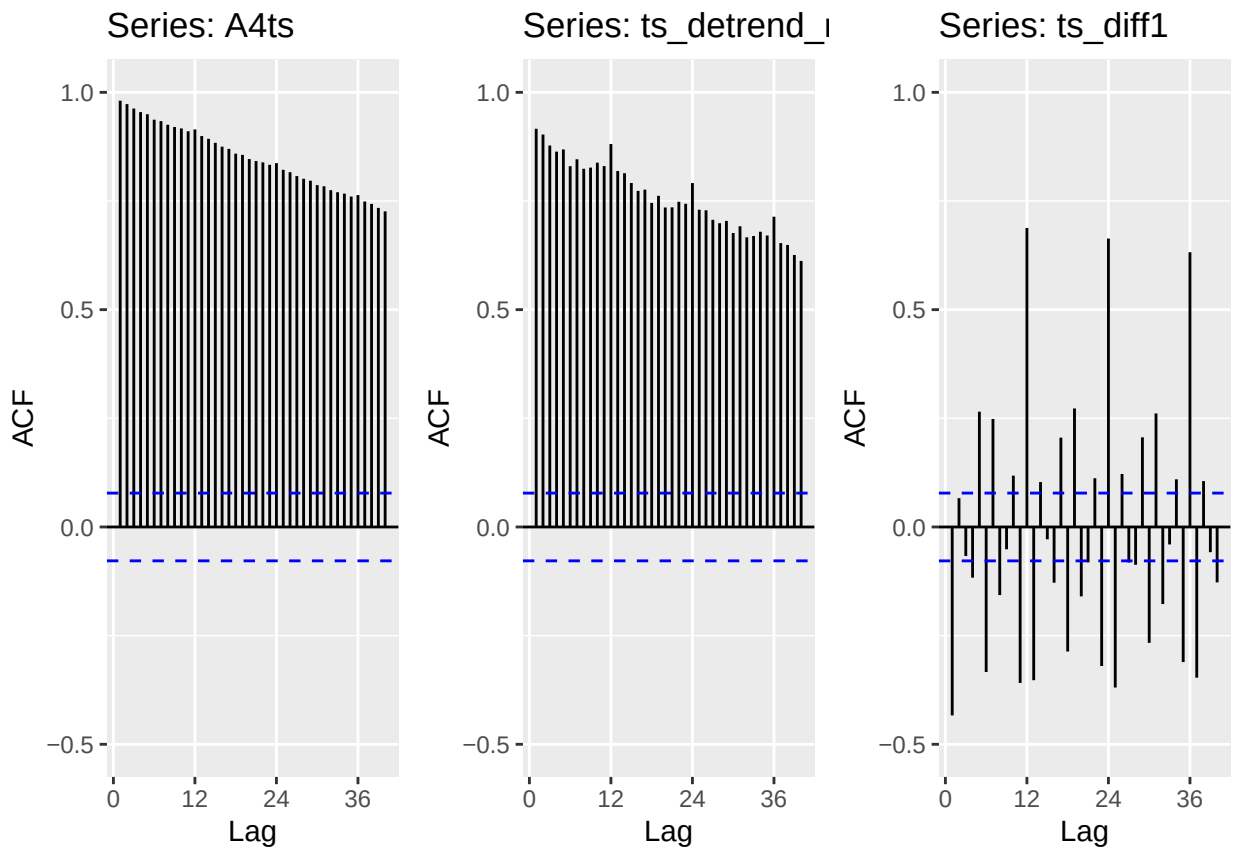

```
## Warning in ggplot2::geom_segment(lineend = "butt", ...): Ignoring unknown
## parameters: `ylim`
```

```
Renewables3 <- autoplot(Acf(ts_diff1, lag.max = 40),ylim =c(-0.5,1))
```




```
## Warning in ggplot2::geom_segment(lineend = "butt", ...): Ignoring unknown
## parameters: `ylim`
```

```
cowplot::plot_grid(Renewables1, Renewables2, Renewables3, nrow = 1)
```



Answer: The linear regression appears to have removed only a minute fraction of the trend, such that we are able to see that there is some seasonality, but we cannot identify the magnitude of the seasonality. On the other hand, the differencing has resulted in a series that has nearly no trend, and instead appears to largely have only seasonal peaks at regular intervals (the 12 month peak is surrounded by 2 strong negative peaks at the 11 and 13th month, etc.)

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 differencing to remove the trend.

```
print(SeasonalMannKendall(A4ts))
```

```
## tau = 0.799, 2-sided pvalue =< 2.22e-16
```

```
print(adf.test(A4ts))
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  A4ts
## Dickey-Fuller = -1.0247, Lag order = 8, p-value = 0.9347
## alternative hypothesis: stationary
```

Answer: The Seasonal Mann-Kendall has a tau value of 0.799, and a 2-sided P-value of much less

than 0.05. There is sufficient evidence to reject the null hypothesis of stationarity and retain the alternate hypothesis that the series follows a trend. The Augemnted Dickey-Fuller test returned a -1.0247 metric, which corresponds to a p-value of 0.9347. As a result, we retain the null hypothesis that the series contains a unit root and thus is non-stationary.

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().

```
# Aggregate by year and month
energy_matrix <- matrix(A4ts,byrow=TRUE,nrow=12)
```

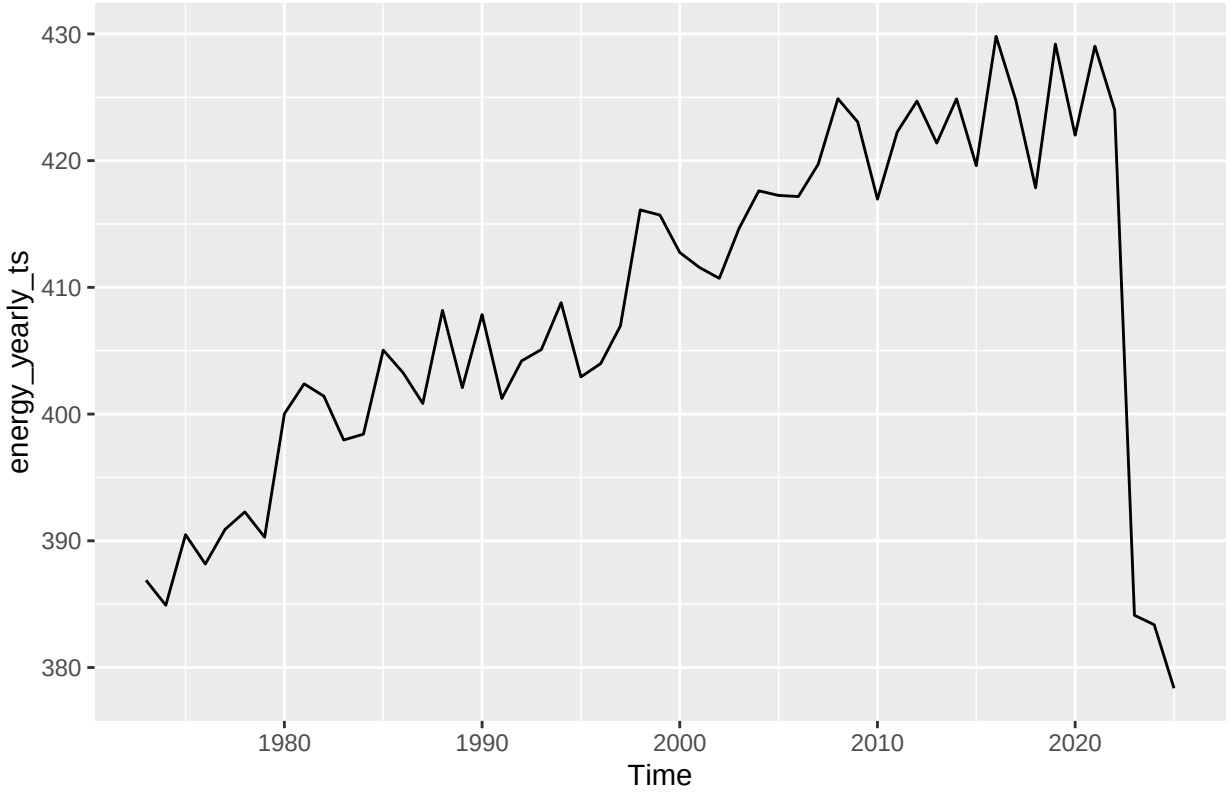
```
## Warning in matrix(A4ts, byrow = TRUE, nrow = 12): data length [633] is not a
## sub-multiple or multiple of the number of rows [12]
```

```
energy_yearly <- colMeans(energy_matrix)

# convert to time series
energy_yearly_ts <- ts(energy_yearly, start= c(1973, 1), frequency = 1)

y <- c(seq(length(energy_yearly)))

#plotlength()#plot
autoplot(energy_yearly_ts)
```



Q7

Apply the Mann Kendall, 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 Q5?

```
MannKendall(energy_yearly_ts)
```

```
## tau = 0.607, 2-sided pvalue =< 2.22e-16
```

```
cor.test(x = energy_yearly_ts, y = y, method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  energy_yearly_ts and y
## S = 8958, p-value = 5.581e-07
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.6388486
```

```
adf.test(energy_yearly_ts)
```

```
## Warning in adf.test(energy_yearly_ts): p-value greater than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  energy_yearly_ts
## Dickey-Fuller = 0.76357, Lag order = 3, p-value = 0.99
## alternative hypothesis: stationary
```

> Answer: the MannKendall test returned very similar results, but dropped slightly from 0.799 to 0.607, while retaining a very statistically significant p-value; this reaffirms the results from Q5, and we can again reject the null hypothesis, and retain the alternate hypothesis that there is a trend. The Spearman's Correlation Rank results demonstrate that there is a very statistically significant monotonic relationship, with a rho of 0.6388. The Augmented Dickey Fuller test returned similar results to the test ran in Question 5: the statistic has a very high p-value, and so we do not reject the null hypothesis that there is nonstationarity.