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

Assignment 3 - Due date 02/01/24

Jaimie Wargo

Directions

You should open the .rmd file corresponding to this assignment on RStudio. The file is available on our class repository on Github.

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_A03_Sp24.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).

Please keep this R code chunk options for the report. It is easier for us to grade when we can see code and output together. And the tidy.opts will make sure that line breaks on your code chunks are automatically added for better visualization.

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

Questions

Consider the same data you used for A2 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 December 2022 **Monthly** Energy Review. Once again you will work only with the following columns: Total Biomass Energy Production, Total Renewable Energy Production, Hydroelectric Power Consumption. Create a data frame structure with these three time series only.

R packages needed for this assignment:"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)
```

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

```
library(tseries)
library(Kendall)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```r
library(here)
```

```
## here() starts at C:/Users/jaimi/OneDrive/Documents/Duke/Spring_2024/TSA_Sp24
```

```r
library(cowplot)
```

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

```r
library(readxl)
```

Trend Component

Q1

For each time series, i.e., Renewable Energy Production and Hydroelectric Consumption create three plots: one with time series, one with the ACF and with the PACF. You may use the same code from A2, but I want all the three plots side by side as in a grid. (Hint: use function `plot_grid()` from the `cowplot` package)

```r
#Importing data set
raw_data <- read_excel(path=here('Data',
          'Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx'),
          skip = 12, sheet="Monthly Data",col_names=FALSE)
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
```

```
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
```

```
#Extract the column names from row 11
read_col_names <- read_excel(path=here('Data',
          'Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx'),
            skip = 10,n_max = 1, sheet="Monthly Data",col_names=FALSE)
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
```

```
colnames(raw_data) <- read_col_names

#Data frame
energy_df <- data.frame(raw_data[,c('Month', 'Total Biomass Energy Consumption',
                         'Total Renewable Energy Consumption',
                         'Hydroelectric Power Consumption')])

nospace_colnames <- c('date', 'biomass', 'total_renew', 'hydro')
colnames(energy_df) <- nospace_colnames

#Time series structure
energy_ts <- ts(energy_df[,2:4], start=c(1973, 1), frequency = 12)
```

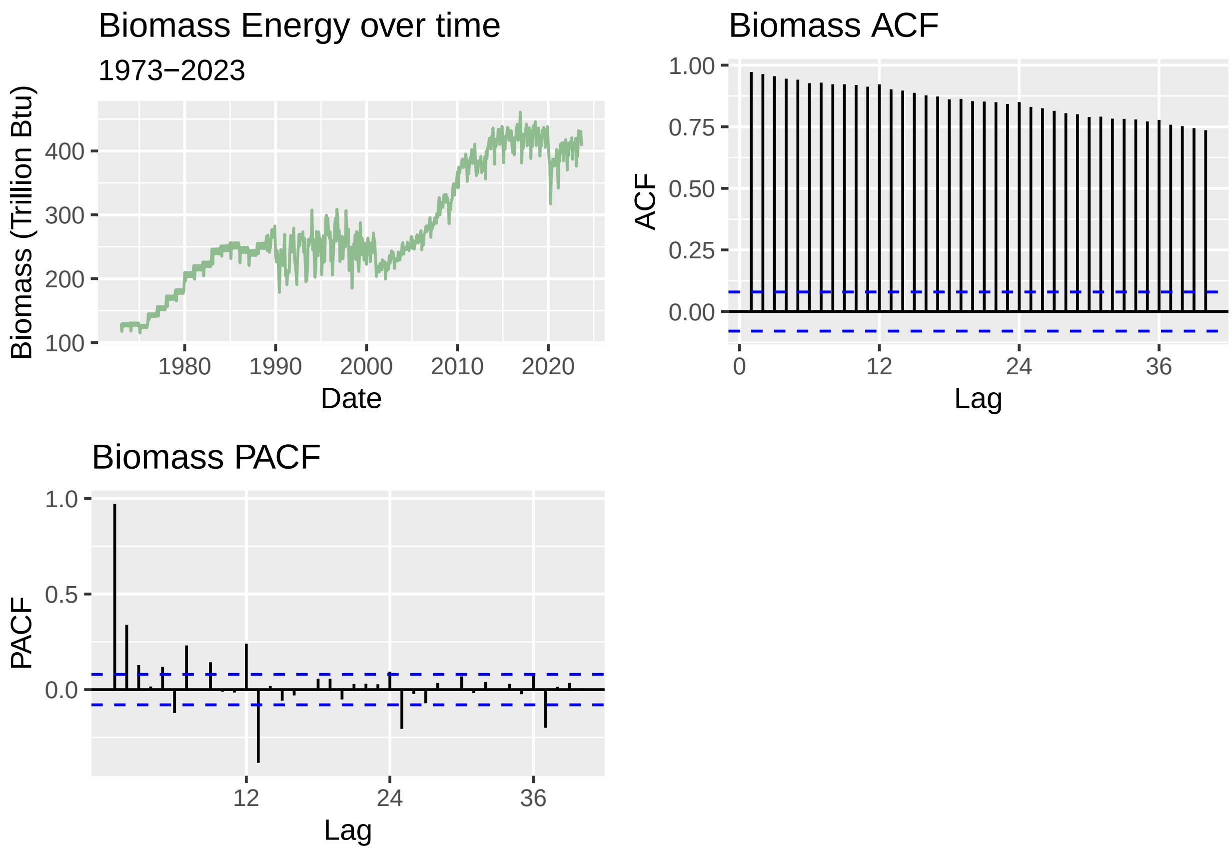
```
bm_tsplot <- ggplot(data=energy_df, aes(x=date, y=biomass))+
  geom_line(color='darkseagreen')+
  labs(x='Date', y='Biomass (Trillion Btu)',
       title='Biomass Energy over time', subtitle = '1973-2023')

bm_acfplot <- autoplot(Acf(energy_ts[,1], lag.max=40, plot=F), main="Biomass ACF")
```

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

```r
bm_pacfplot <- autoplot(Pacf(energy_ts[,1], lag.max=40, plot=F), main="Biomass PACF")
```

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

```r
plot_grid(bm_tsplot, bm_acfplot, bm_pacfplot)
```



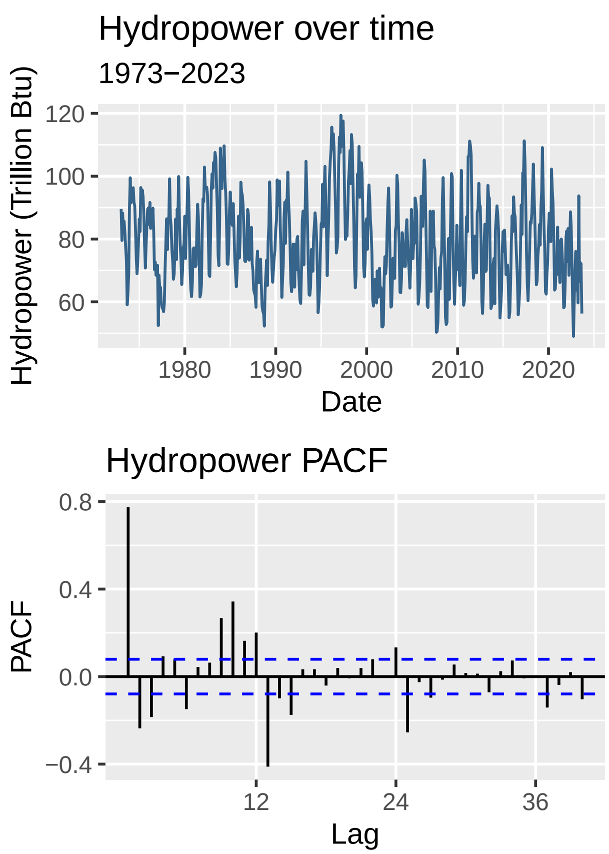

```r
hp_tsplot <- ggplot(data=energy_df, aes(x=date, y=hydro))+
  geom_line(color='steelblue4')+
  labs(x='Date', y='Hydropower (Trillion Btu)',
       title='Hydropower over time', subtitle = '1973-2023')

hp_acfplot <- autoplot(Acf(energy_ts[,3], lag.max=40, plot=F), main="Hydropower ACF")
```

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

```r
hp_pacfplot <- autoplot(Pacf(energy_ts[,3], lag.max=40, plot=F), main="Hydropower PACF")
```

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

```r
plot_grid(hp_tsplot, hp_acfplot, hp_pacfplot)
```



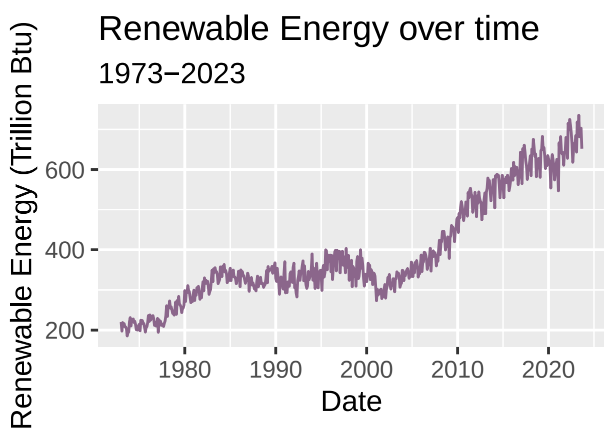

```r
tre_tsplot <- ggplot(data=energy_df, aes(x=date, y=total_renew))+
  geom_line(color='plum4')+
  labs(x='Date', y='Renewable Energy (Trillion Btu)',
       title='Renewable Energy over time', subtitle = '1973-2023')

tre_acfplot <- autoplot(Acf(energy_ts[,2], lag.max=40, plot=F),main="Renewable Energy ACF")
```

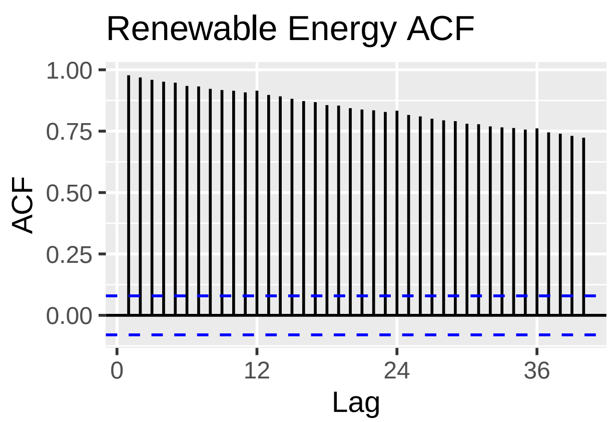
```
## Warning in ggplot2::geom_segment(lineend = "butt", ...): Ignoring unknown
## parameters: ‘main‘
```

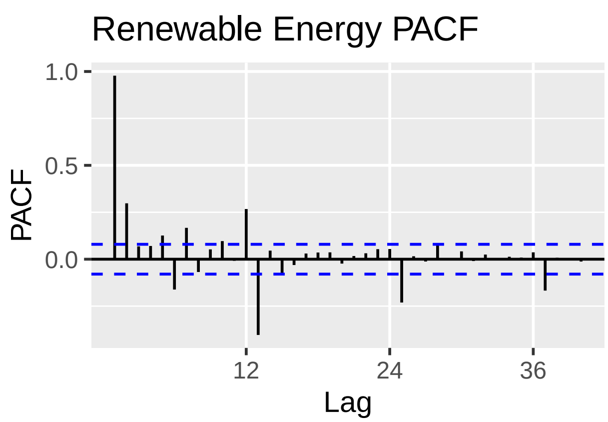
```r
tre_pacfplot <- autoplot(Pacf(energy_ts[,2], lag.max=40,plot=F),main="Renewable Energy PACF")
```

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

```r
plot_grid(tre_tsplot, tre_acfplot, tre_pacfplot)
```







Q2

From the plot in Q1, do the series Total Biomass Energy Production, Total Renewable Energy Production, Hydroelectric Power Consumption appear to have a trend? If yes, what kind of trend?

Answer: The biomass and total renewable energy plots both seem to have an upwards, roughly linear trend, whereas hydroelectric appears constant over the course of the observations.

Q3

Use the *lm()* function to fit a linear trend to the three time series (I am going to just use total renewable and hydro as indicated in lecture 2/1). Ask R to print the summary of the regression. Interpret the regression output, i.e., slope and intercept. Save the regression coefficients for further analysis.

```
nobs <- nrow(energy_df)
t <- 1:nobs

renewable_trend <- lm(energy_df[,3]~t)
summary(renewable_trend)
```

```
##
## Call:
## lm(formula = energy_df[, 3] ~ t)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -144.72  -33.24   11.30   39.21  135.99
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 184.40303    4.69129   39.31   <2e-16 ***
## t             0.68542    0.01333   51.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 57.81 on 607 degrees of freedom
## Multiple R-squared:  0.8134, Adjusted R-squared:  0.8131
## F-statistic:  2645 on 1 and 607 DF,  p-value: < 2.2e-16
```

```r
tre_beta0 <- renewable_trend$coefficients[1]
tre_beta1 <- renewable_trend$coefficients[2]
```

The R2 of 0.8131 indicates that 81.31% of variation can be explained by the trend. The overall slope is estimated at 0.685, indicating an increasing trend, and the intercept of 184 indicates the starting value as modeled linearly, starting at 184 trillion Btu of renewable energy in 1973.

```r
hydro_trend <- lm(energy_df[,4]~t)
summary(hydro_trend)
```

```
##
## Call:
## lm(formula = energy_df[, 4] ~ t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.818 -10.620  -0.669   9.357  39.528
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 82.734747   1.140265  72.557  < 2e-16 ***
## t           -0.009849   0.003239  -3.041  0.00246 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.05 on 607 degrees of freedom
## Multiple R-squared:  0.015,  Adjusted R-squared:  0.01338
## F-statistic: 9.247 on 1 and 607 DF,  p-value: 0.002461
```

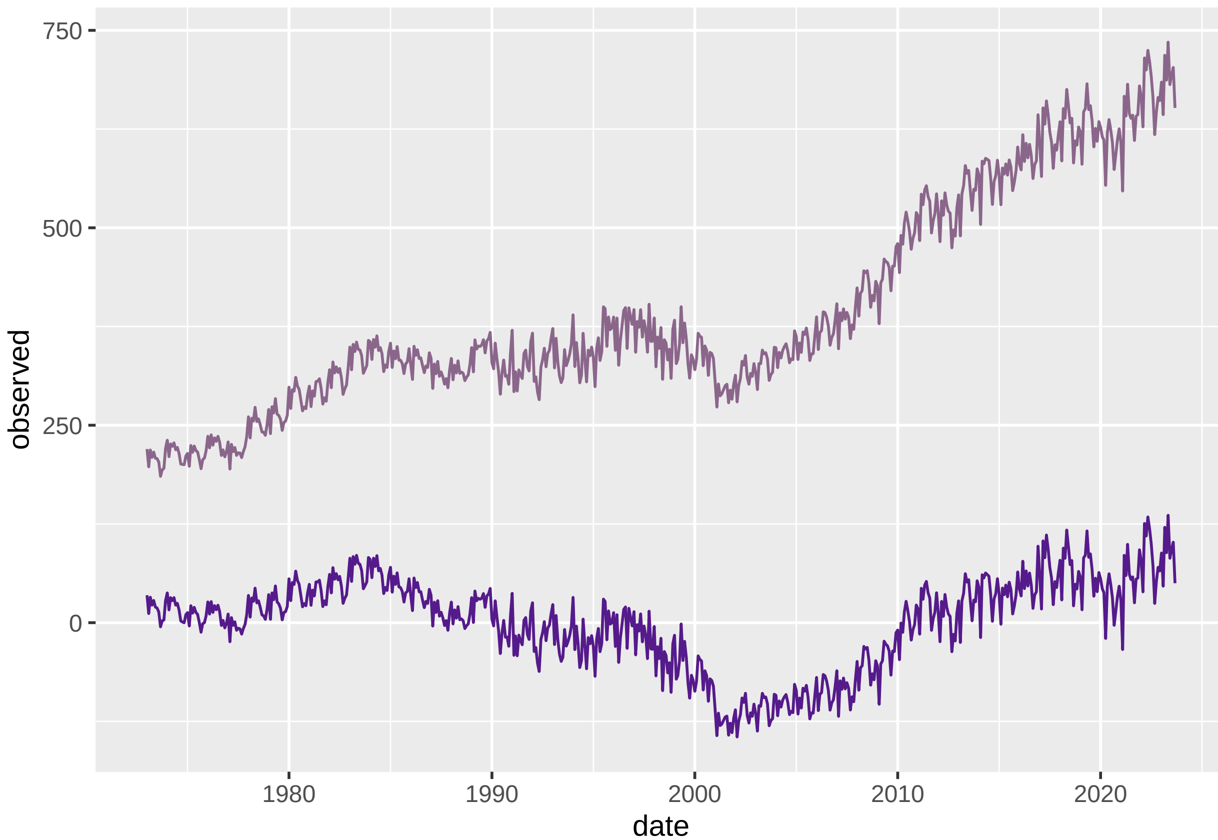
```r
hp_beta0 <- hydro_trend$coefficients[1]
hp_beta1 <- hydro_trend$coefficients[2]
```

The R2 of 0.013 indicates that 1.3% of variation can be explained by the trend. The overall slope is estimated at -0.00985, indicating a fairly flat trend (or lack thereof), and the intercept of 82.73 indicates the starting value as modeled linearly, starting at 82.73 trillion Btu of hydroelectric energy in 1973.

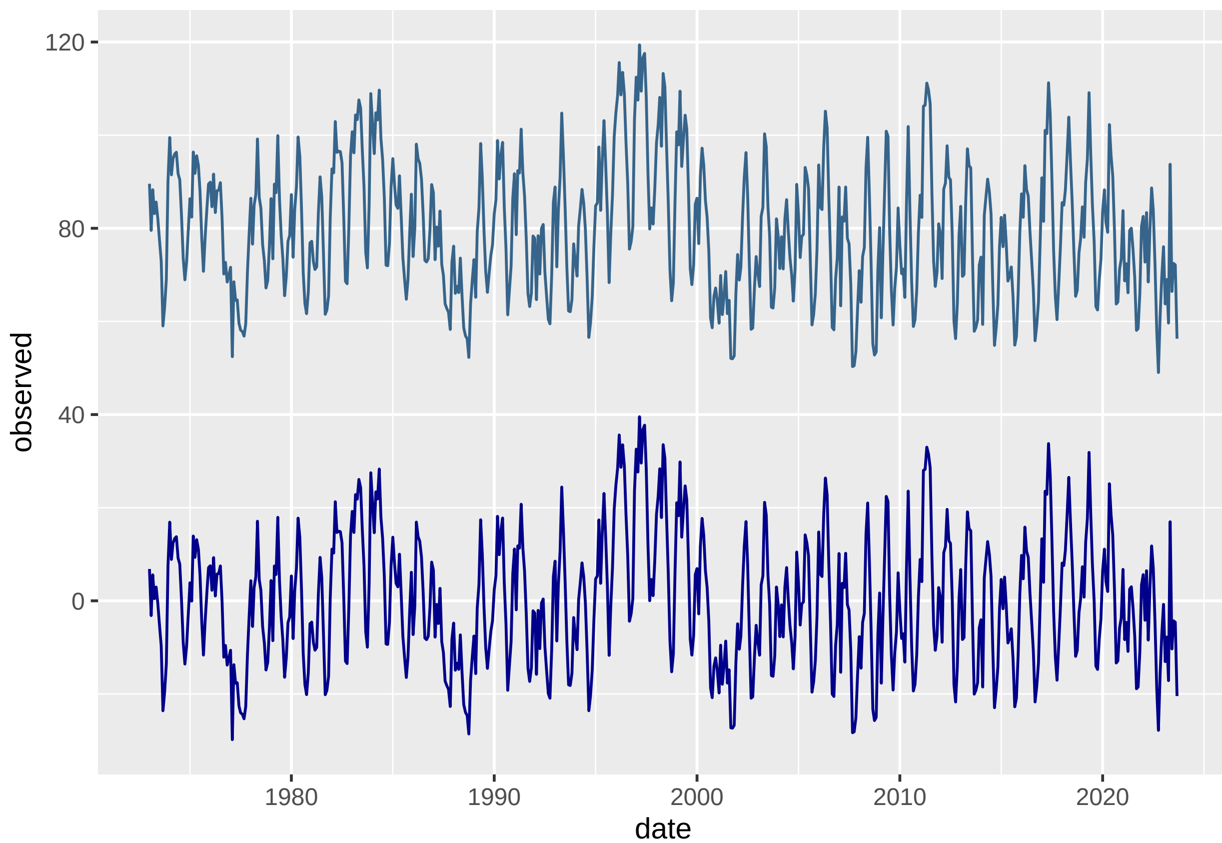
Q4

Use the regression coefficients from Q3 to detrend the series. Plot the detrended series and compare with the plots from Q1. What happened? Did anything change?

```
tre_detrend <- energy_df[,3] - (tre_beta0+tre_beta1*t)

tre_df <- data.frame('date'=energy_df$date,
                         'observed'=energy_df[,3],
                         'detrend'=tre_detrend)

ggplot(tre_df, aes(x=date))+
  geom_line(aes(y=observed), color='plum4')+
  geom_line(aes(y=detrend), color='purple4')
```




```
hp_detrend <- energy_df[,4] - (hp_beta0+hp_beta1*t)

hp_df <- data.frame('date'=energy_df$date,
                        'observed'=energy_df[,4],
                        'detrend'=hp_detrend)

ggplot(hp_df, aes(x=date))+
  geom_line(aes(y=observed), color='steelblue4')+
  geom_line(aes(y=detrend), color='darkblue')
```



The renewable energy plot flattened significantly as it lost its upward trend. The hydro plot just shifted down to center at 0, but the shape of the series did not change. This makes sense because the trend of the renewables plot was very influential to the plot's shape, whereas the trend for the hydro series was not.

Q5

Plot ACF and PACF for the detrended series and compare with the plots from Q1. You may use plot_grid() again to get them side by side. but not mandatory. Did the plots change? How?

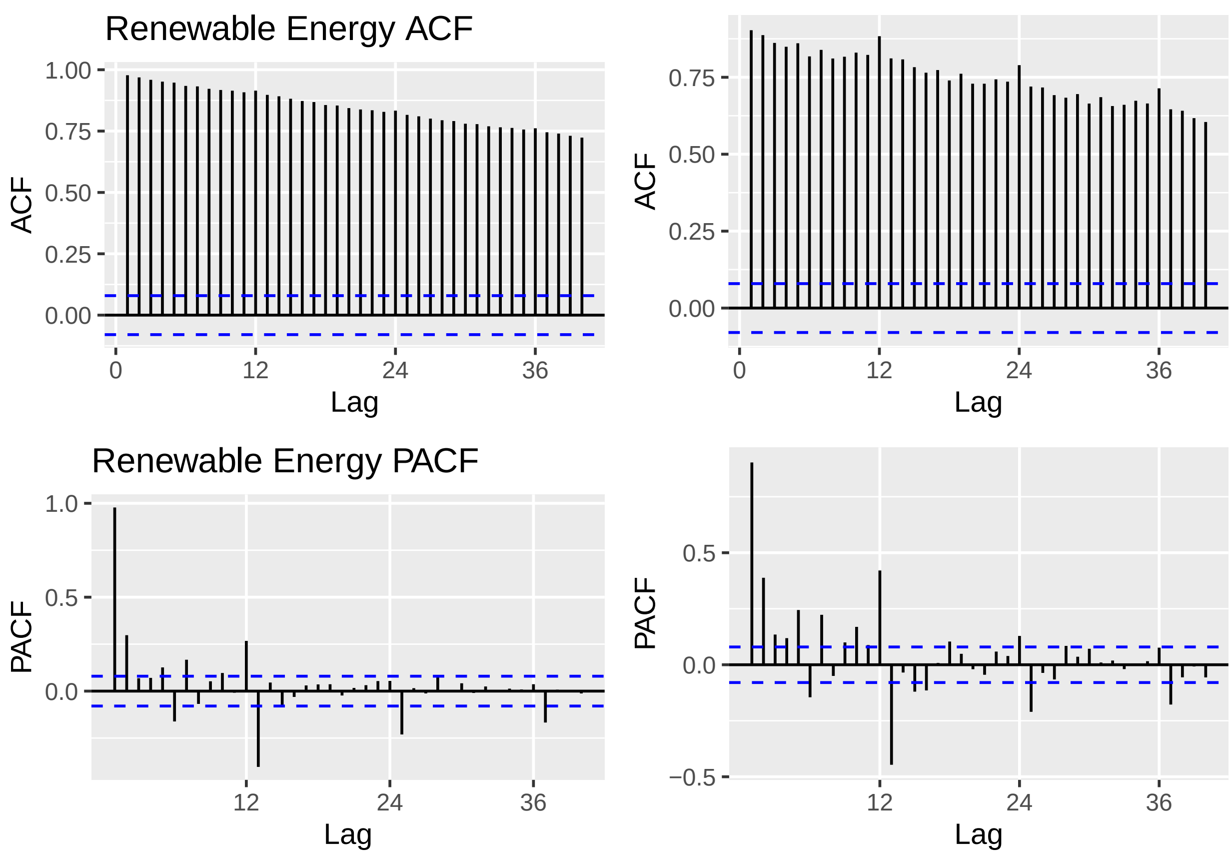
```
tre_ts_detrend <- ts(tre_df$detrend, frequency = 12, start(1973,1))

tre_acfplot_dt <- autoplot(Acf(tre_ts_detrend, lag.max=40, plot=F), main=NULL)
```

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

```
tre_pacfplot_dt <- autoplot(Pacf(tre_ts_detrend, lag.max=40,plot=F),main=NULL)
```

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

```
plot_grid(tre_acfplot, tre_acfplot_dt, tre_pacfplot, tre_pacfplot_dt)
```



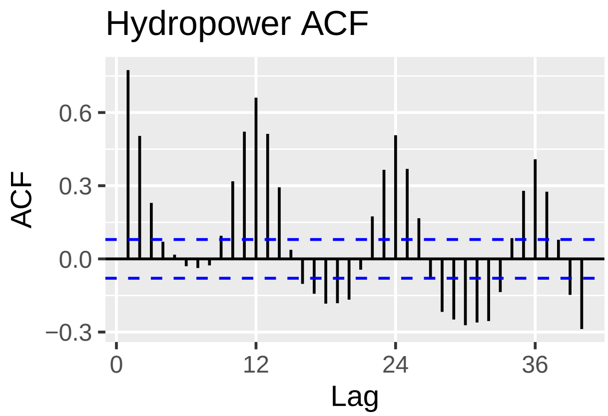

```r
hp_ts_detrend <- ts(hp_df$detrend, frequency = 12, start(1973,1))

hp_acfplot_dt <- autoplot(Acf(hp_ts_detrend, lag.max=40, plot=F), main=NULL)
```

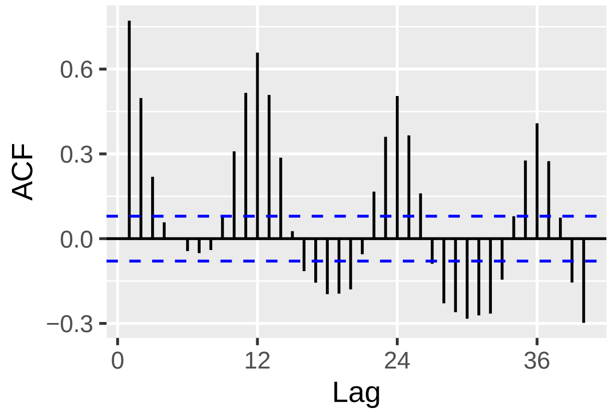
```
## Warning in ggplot2::geom_segment(lineend = "butt", ...): Ignoring unknown
## parameters: `main`
```

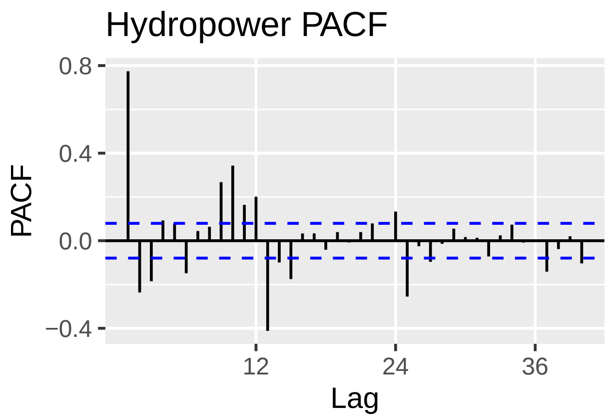
```r
hp_pacfplot_dt <- autoplot(Pacf(hp_ts_detrend, lag.max=40,plot=F),main=NULL)
```

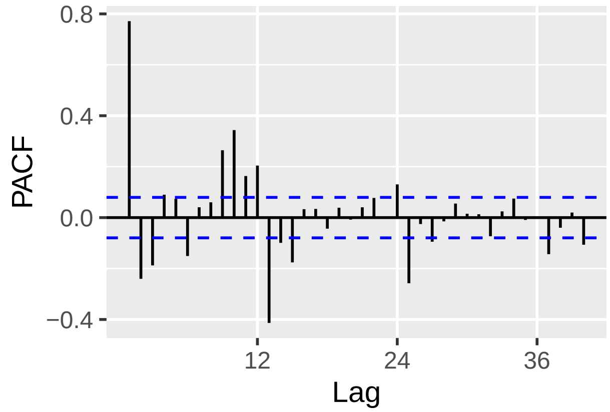
```
## Warning in ggplot2::geom_segment(lineend = "butt", ...): Ignoring unknown
## parameters: `main`
```

```r
plot_grid(hp_acfplot, hp_acfplot_dt, hp_pacfplot, hp_pacfplot_dt)
```









The ACF and PACF plots did not appear to change from the removal of the trend. There were a few more prominent lags for the renewables PACF, with lags 12 and 13 having PACFs further from zero, but there were no notable shifts in these plots.

Seasonal Component

Set aside the detrended series and consider the original series again from Q1 to answer Q6 to Q8.

Q6

Just by looking at the time series and the acf plots, do the series seem to have a seasonal trend? No need to run any code to answer your question. Just type in you answer below.

Answer: The hydro plots strongly suggestthat there is a seasonal trend, which makes sense since inflows are dependent on weather. The renewables plots do not have an obvious seasonal trend component just from visual assessment.

Q7

Use function *lm()* to fit a seasonal means model (i.e. using the seasonal dummies) the two time series. Ask R to print the summary of the regression. Interpret the regression output. From the results which series have a seasonal trend? Do the results match you answer to Q6?

```r
dummies <- seasonaldummy(energy_ts[,2])

tre_seas_means_model <- lm(energy_df[,3]~dummies)
summary(tre_seas_means_model)
```

```
##
## Call:
## lm(formula = energy_df[, 3] ~ dummies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -197.20  -83.65  -46.46  112.68  325.04
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 401.6658    18.9908  21.151   <2e-16 ***
## dummiesJan    0.7032    26.7251   0.026    0.979
## dummiesFeb  -34.7418    26.7251  -1.300    0.194
## dummiesMar    3.3631    26.7251   0.126    0.900
## dummiesApr   -7.9593    26.7251  -0.298    0.766
## dummiesMay    8.3640    26.7251   0.313    0.754
## dummiesJun   -2.4081    26.7251  -0.090    0.928
## dummiesJul    3.3382    26.7251   0.125    0.901
## dummiesAug   -3.1366    26.7251  -0.117    0.907
## dummiesSep  -28.0412    26.7251  -1.049    0.294
## dummiesOct  -18.3929    26.8571  -0.685    0.494
## dummiesNov  -19.8835    26.8571  -0.740    0.459
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 134.3 on 597 degrees of freedom
## Multiple R-squared:  0.009755,   Adjusted R-squared:  -0.008491
## F-statistic: 0.5347 on 11 and 597 DF,  p-value: 0.8802
```

```r
tre_beta_int <- tre_seas_means_model$coefficients[1]
tre_beta_coeff <- tre_seas_means_model$coefficients[2:12]
```

```r
dummies <- seasonaldummy(energy_ts[,3])

hp_seas_means_model <- lm(energy_df[,4]~dummies)
summary(hp_seas_means_model)
```

```
##
## Call:
## lm(formula = energy_df[, 4] ~ dummies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.323  -5.849  -0.468   6.243  32.290
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   80.282      1.470  54.601  < 2e-16 ***
## dummiesJan     4.807      2.069   2.323  0.02050 *
## dummiesFeb    -2.725      2.069  -1.317  0.18831
## dummiesMar     6.825      2.069   3.298  0.00103 **
## dummiesApr     5.319      2.069   2.571  0.01039 *
## dummiesMay    13.922      2.069   6.729 4.02e-11 ***
## dummiesJun    10.650      2.069   5.147 3.60e-07 ***
## dummiesJul     3.912      2.069   1.891  0.05914 .
## dummiesAug    -5.677      2.069  -2.744  0.00626 **
## dummiesSep   -16.797      2.069  -8.118 2.72e-15 ***
## dummiesOct   -16.468      2.079  -7.920 1.17e-14 ***
## dummiesNov   -10.885      2.079  -5.235 2.29e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.4 on 597 degrees of freedom
## Multiple R-squared:  0.4697, Adjusted R-squared:  0.4599
## F-statistic: 48.07 on 11 and 597 DF,  p-value: < 2.2e-16
```

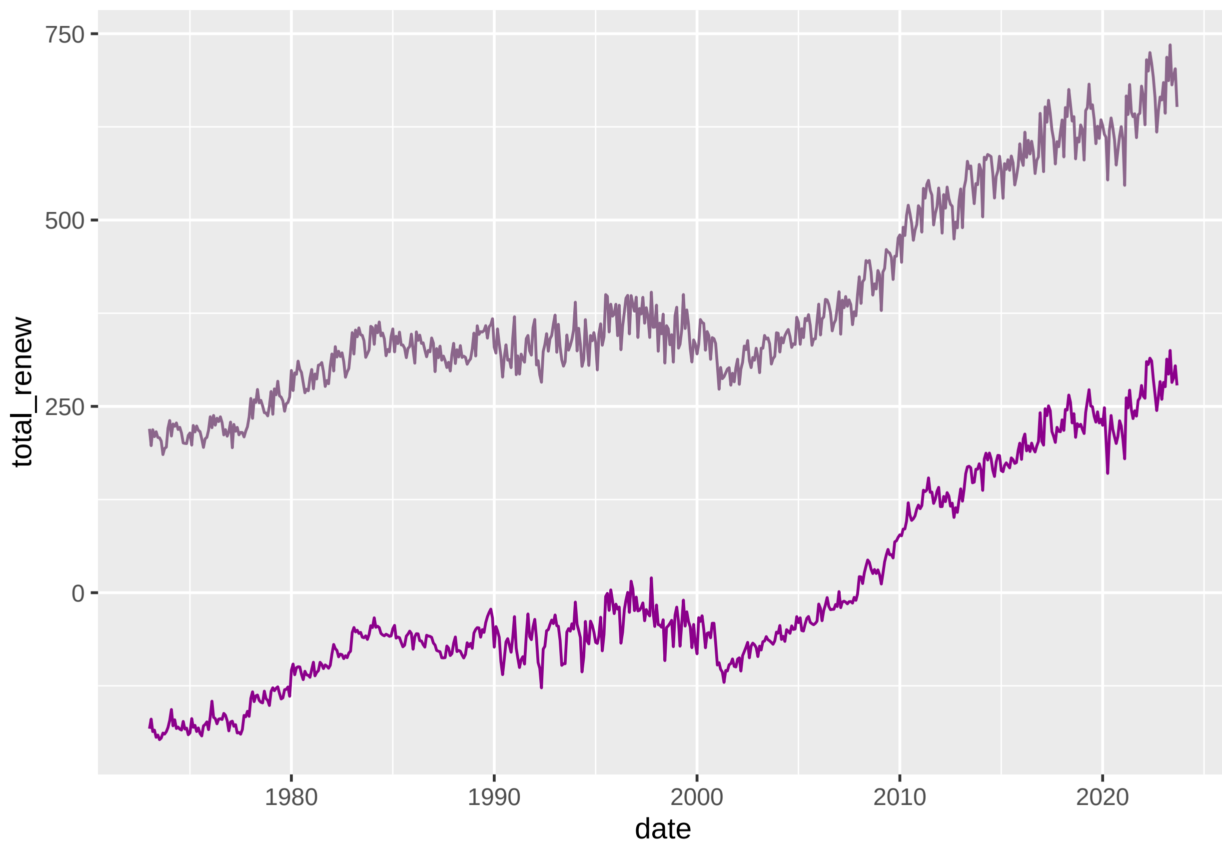
```
hp_beta_int <- hp_seas_means_model$coefficients[1]
hp_beta_coeff <- hp_seas_means_model$coefficients[2:12]
```

The results do match my assessment in Q6. The hydro seasonality explains 45.99% of variation in the model, while the renewables seasonal model explain 0.9% (unadjusted R2). This means that hydro does have a seasonal component, while total renewable energy does not.

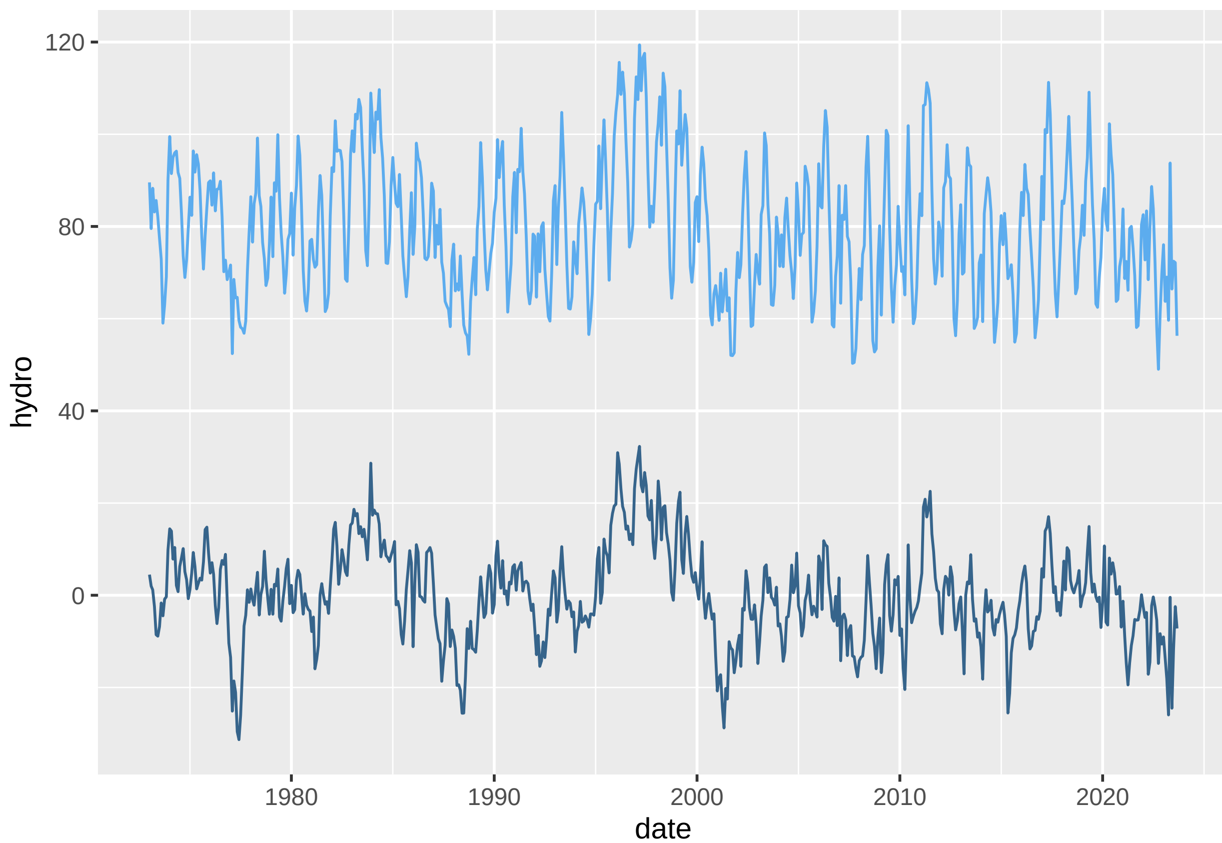
Q8

Use the regression coefficients from Q7 to deseason the series. Plot the deseason series and compare with the plots from part Q1. Did anything change?

```
tre_seas_comp <- array(0,nobs)

for(i in 1:nobs){
  tre_seas_comp[i] <- (tre_beta_int+tre_beta_coeff %*% dummies[i,])
}

deseason_tre_data <- energy_df[,3]-tre_seas_comp

ggplot(energy_df, aes(x=date, y=total_renew)) +
            geom_line(color="plum4") +
            geom_line(aes(y=deseason_tre_data), col="darkmagenta")
```




```
hp_seas_comp <- array(0,nobs)

for(i in 1:nobs){
  hp_seas_comp[i] <- (hp_beta_int+hp_beta_coeff %*% dummies[i,])
}

deseason_hp_data <- energy_df[,4]-hp_seas_comp

ggplot(energy_df, aes(x=date, y=hydro)) +
            geom_line(color="steelblue2") +
            geom_line(aes(y=deseason_hp_data), col="steelblue4")
```



The renewables plot just shifted down, but the shape was not impacted by removing the seasonal component. For the hydro series, there is a significant change in the shape. This shows that the hydro series did in fact have a strong seasonal component.

Q9

Plot ACF and PACF for the deseason series and compare with the plots from Q1. You may use plot_grid() again to get them side by side. but not mandatory. Did the plots change? How?

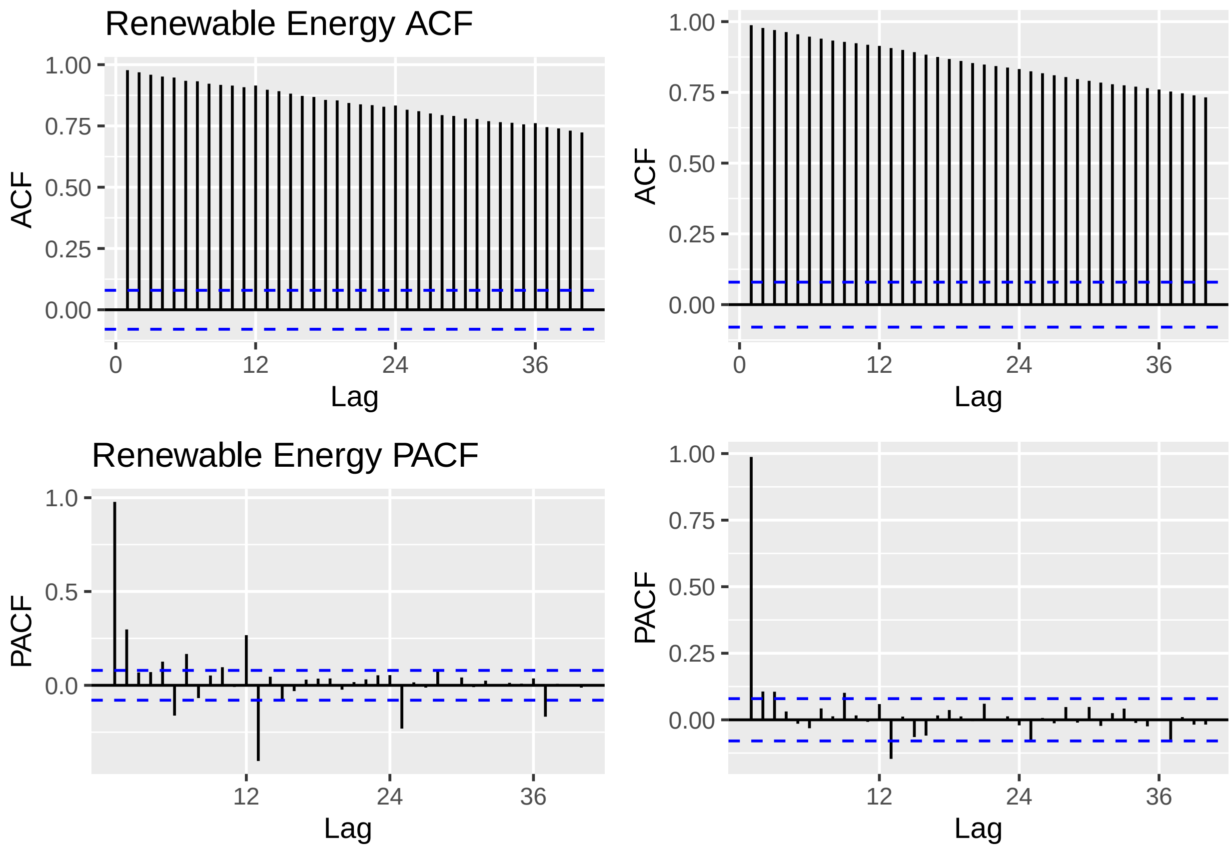
```
tre_ts_deseason<- ts(deseason_tre_data, frequency = 12, start(1973,1))

tre_acfplot_ds <- autoplot(Acf(tre_ts_deseason, lag.max=40, plot=F), main=NULL)
```

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

```
tre_pacfplot_ds <- autoplot(Pacf(tre_ts_deseason, lag.max=40,plot=F),main=NULL)
```

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

```
plot_grid(tre_acfplot, tre_acfplot_ds, tre_pacfplot, tre_pacfplot_ds)
```



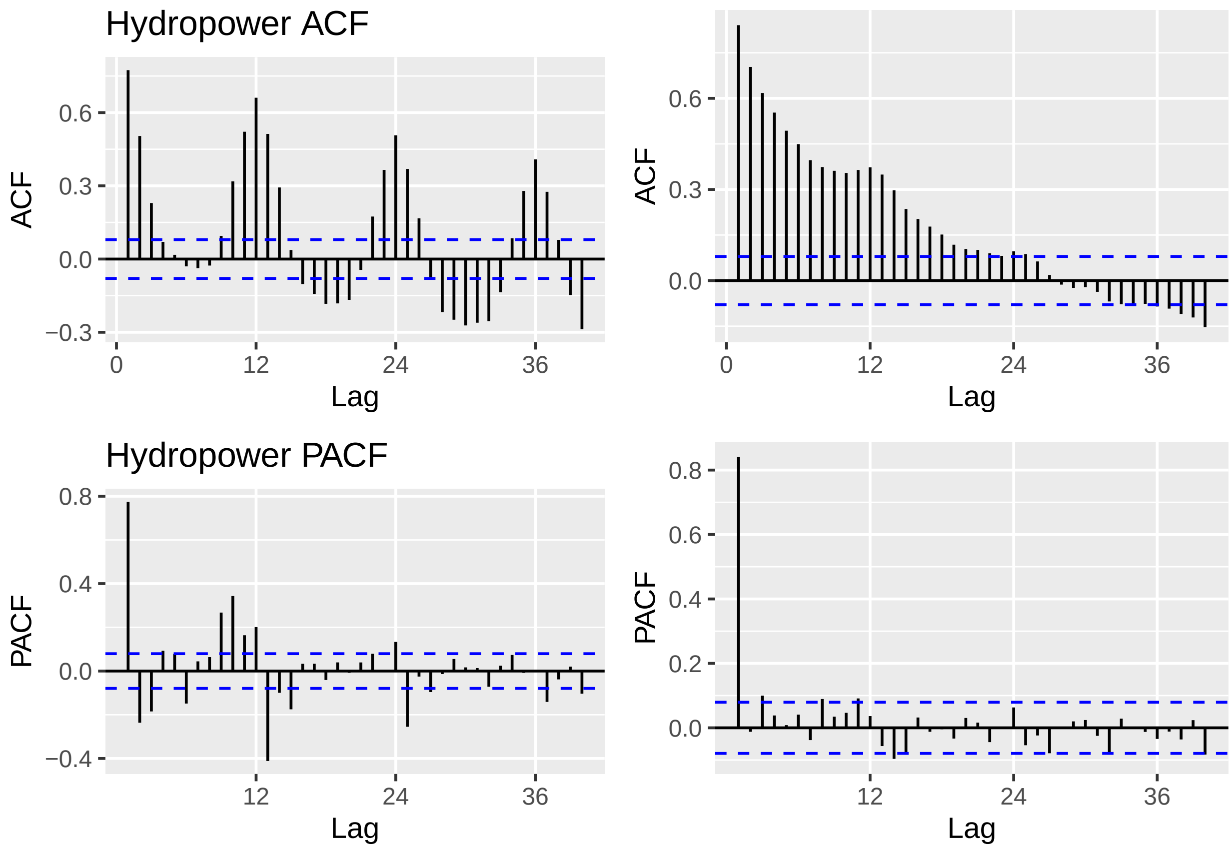

```r
hp_ts_deseason <- ts(deseason_hp_data, frequency = 12, start(1973,1))

hp_acfplot_ds <- autoplot(Acf(hp_ts_deseason, lag.max=40, plot=F), main=NULL)
```

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

```r
hp_pacfplot_ds <- autoplot(Pacf(hp_ts_deseason, lag.max=40,plot=F),main=NULL)
```

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

```r
plot_grid(hp_acfplot, hp_acfplot_ds, hp_pacfplot, hp_pacfplot_ds)
```



The renewable energy plots again do not change much. However, the hydropower plots have a significant difference after being deseasoned. The ACF changes from oscillating to having slowly declining values for each lag, and the PACF mimics this with a strong lag 1 value and no other strong lags.