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

Assignment 3 - Due date 02/01/24

Zhenghao Lin

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_A02_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(lubridate)
```

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

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

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

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

```r
library(Kendall)
library(tseries)
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
```

```r
library(cowplot)
```

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

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

##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 some code form 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
getwd()
```

```
## [1] "/Users/lzh/Desktop/TSA_Sp24"
```

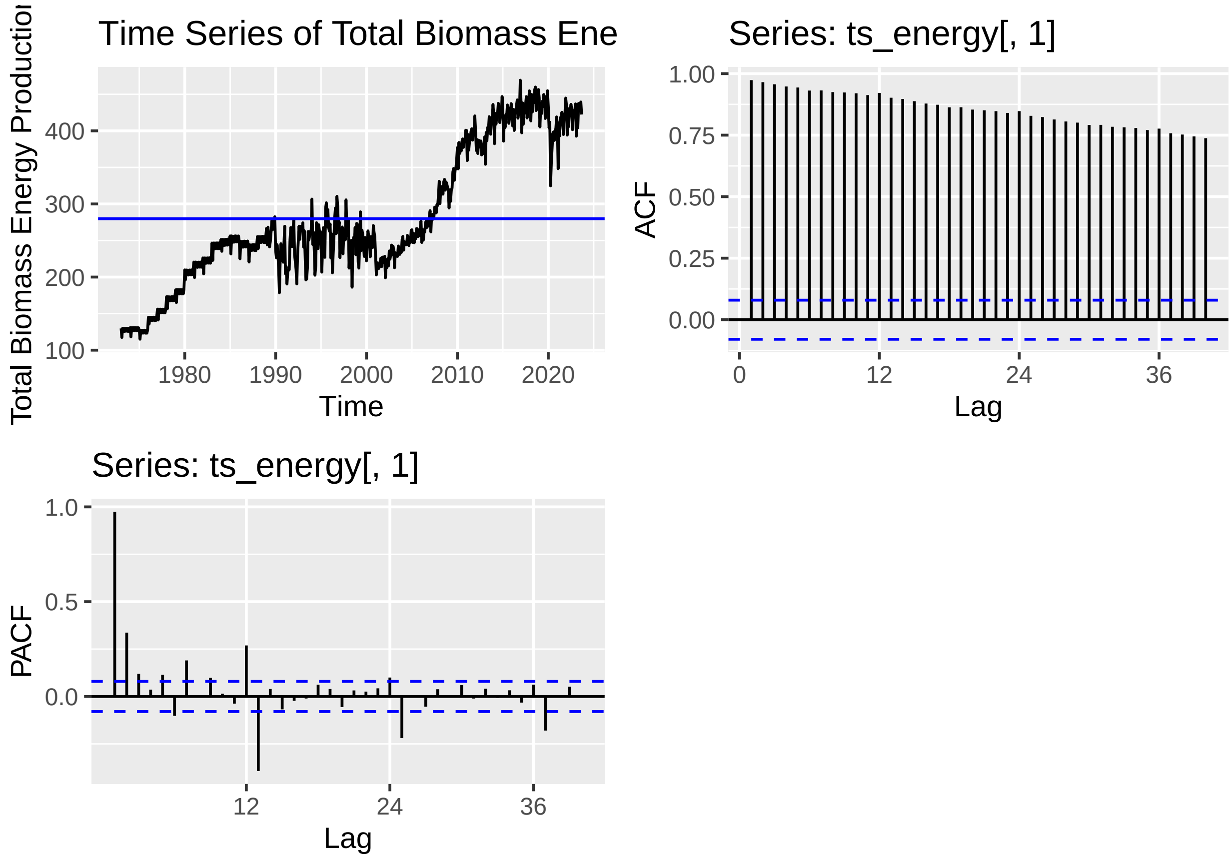
```r
energy_data <- read.table(file="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source

#Date Conversion
ym_date <- ym(energy_data$Month)
```

```r
#New data frame
energy_data <- cbind(ym_date, energy_data[,4:6])

#Verification
head(energy_data)
```

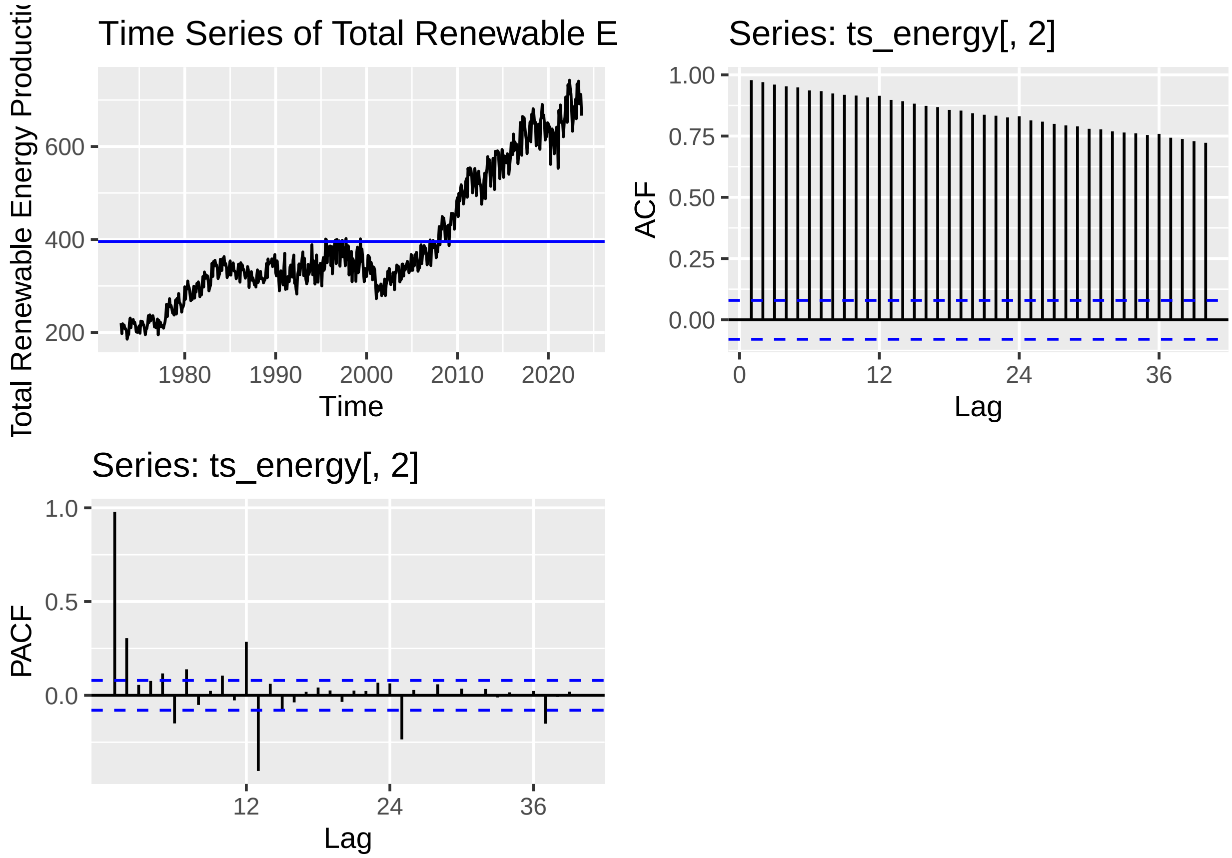
```
##      ym_date Total.Biomass.Energy.Production Total.Renewable.Energy.Production
## 1 1973-01-01                         129.787                           219.839
## 2 1973-02-01                         117.338                           197.330
## 3 1973-03-01                         129.938                           218.686
## 4 1973-04-01                         125.636                           209.330
## 5 1973-05-01                         129.834                           215.982
## 6 1973-06-01                         125.611                           208.249
##   Hydroelectric.Power.Consumption
## 1                          89.562
## 2                          79.544
## 3                          88.284
## 4                          83.152
## 5                          85.643
## 6                          82.060
```

```r
#Columns needed for for loop
nrepc <- ncol(energy_data)

#Time-series data frame
ts_energy <- ts(energy_data[,2:4], start = c(1973, 1), frequency = 12)

#Mean and standard deviation for total biomass energy production
biomass_total_mean <- mean(ts_energy[,1])
biomass_total_stdv <- sd(ts_energy[,1])
#Mean and standard deviation for total renewable energy production
renewable_total_mean <- mean(ts_energy[,2])
renewable_total_stdv <- sd(ts_energy[,2])
#Mean and standard deviation for hydroelectric power consumption
hydro_mean <- mean(ts_energy[,3])
hydro_stdv <- sd(ts_energy[,3])
```

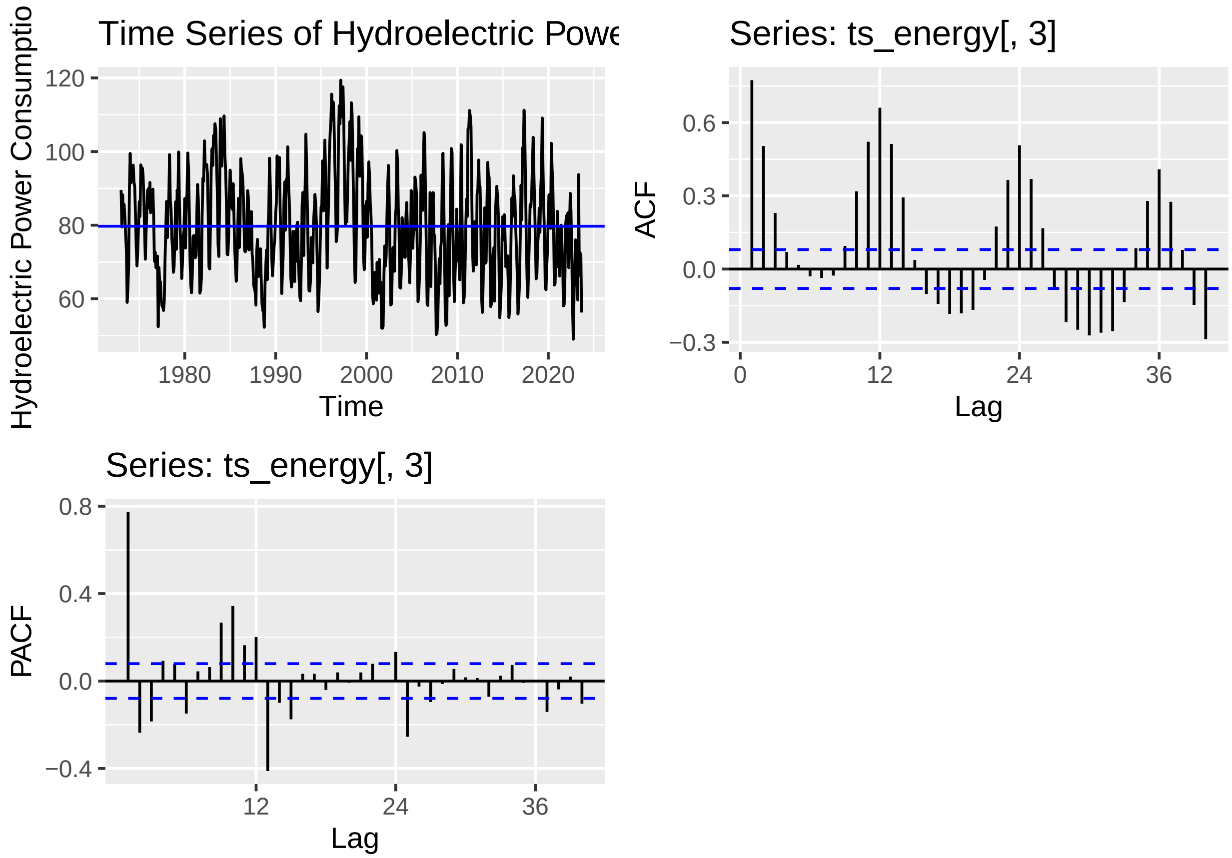
```r
#Plot
#Total Biomass Energy Production Plot Grid
biomass_grid <- plot_grid(biomass_plot, autoplot(biomass_Acf), autoplot(biomass_Pacf))
biomass_grid
```




```r
#Total.Renewable.Energy.Production Plot Grid

rnw_grid <- plot_grid(rnw_plot, autoplot(rnw_Acf), autoplot(rnw_Pacf))
rnw_grid
```




```r
#Hydroelectric Power Consumption Plot Grid
hydro_grid <- plot_grid(hydro_plot, autoplot(hydro_Acf), autoplot(hydro_Pacf))
hydro_grid
```



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: Graphs for biomass energy production and renewable production seem to appear an upward linear trend, yet the trend graph of hydroelectric power consumption is not very clear. There may exist a non-linear trend.

Q3

Use the *lm()* function to fit a linear trend to the three time series. 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.

```r
n <- nrow(energy_data)
t <- 1:n

#Linear trend of Total Biomass Energy Production
trend_biomass <- lm(energy_data[,2]~t)
summary(trend_biomass)
```

```
##
## Call:
```

```
## lm(formula = energy_data[, 2] ~ t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -102.344  -23.754    5.491   31.980   83.154
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 134.27841    3.18601   42.15   <2e-16 ***
## t             0.47713    0.00905   52.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.26 on 607 degrees of freedom
## Multiple R-squared:  0.8208, Adjusted R-squared:  0.8205
## F-statistic:  2780 on 1 and 607 DF,  p-value: < 2.2e-16
```

```
#Coefficients of the linear trend of Total Biomass Energy Production
beta0_bio <- trend_biomass$coefficients[1]
beta1_bio <- trend_biomass$coefficients[2]

#Linear trend of Total Renewable Energy Production
trend_rnw <- lm(energy_data[,3]~t)
summary(trend_rnw)
```

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

```
#Coefficients of the linear trend of Total Renewable Energy Production
beta0_rnw <- trend_rnw$coefficients[1]
beta1_rnw <- trend_rnw$coefficients[2]


#Linear trend of Hydroelectric Power Consumption
trend_hydro <- lm(energy_data[,4]~t)
summary(trend_hydro)
```

```
##
```

```
## Call:
## lm(formula = energy_data[, 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
#Coefficients of the linear trend of Hydroelectric Power Consumption
beta0_hydro <- trend_hydro$coefficients[1]
beta1_hydro <- trend_hydro$coefficients[2]
```

Biomass: Coefficient Intercept (Estimate): The intercept represents the estimated initial level of Total Biomass Energy Production at the beginning of the time series. In this case, the estimated intercept is approximately 134.27841. t (Estimate): The coefficient for the "t" variable represents the estimated rate of change in Total Biomass Energy Production per unit time. In this case, the estimated slope is approximately 0.47713. It means that, on average, Total Biomass Energy Production increases by about 0.47713 units per month.

P-values The p-values in this case are extremely small, which suggests that there is a linear trend in Total Biomass Energy Production.

Model fit The r^2 is about 0.82, which suggests that 82.08% of the variability in Total Biomass Energy Production is accounted for by the linear trend model. The f-stats is very high (2780) with a corresponding p-value close to zero, indicating that the linear model is statistically significant in explaining the variability in Total Biomass Energy Production.

Renewables: Coefficients Intercept (Estimate): The intercept represents the estimated initial level of Total Renewable Energy Production at the beginning of the time series. In this case, the estimated intercept is approximately 184.40303. t (Estimate): The coefficient for the "t" variable represents the estimated rate of change in Total Renewable Energy Production per unit time. In this case, the estimated slope is approximately 0.68542. It means that, on average, Total Renewable Energy Production increases by about 0.68542 units per month.

P-values The p-values in this case are extremely small, which suggests that there is a linear trend in total renewable energy production.

Model fit Both the r^2 and f-stats indicates that the linear model is significant in explaining the variability in total renewable energy production.

Hydroelectric: Coefficients Intercept (Estimate): The intercept represents the estimated initial level of Hydroelectric Power Consumption at the beginning of the time series. In this case, the estimated intercept is approximately 82.734747. t (Estimate): The coefficient for the "t" variable represents the estimated rate of change in Hydroelectric Power Consumption per unit time. In this case, the estimated slope is approximately -0.009849. It means that, on average, Hydroelectric Power Consumption decreases by about 0.009849 units per month.

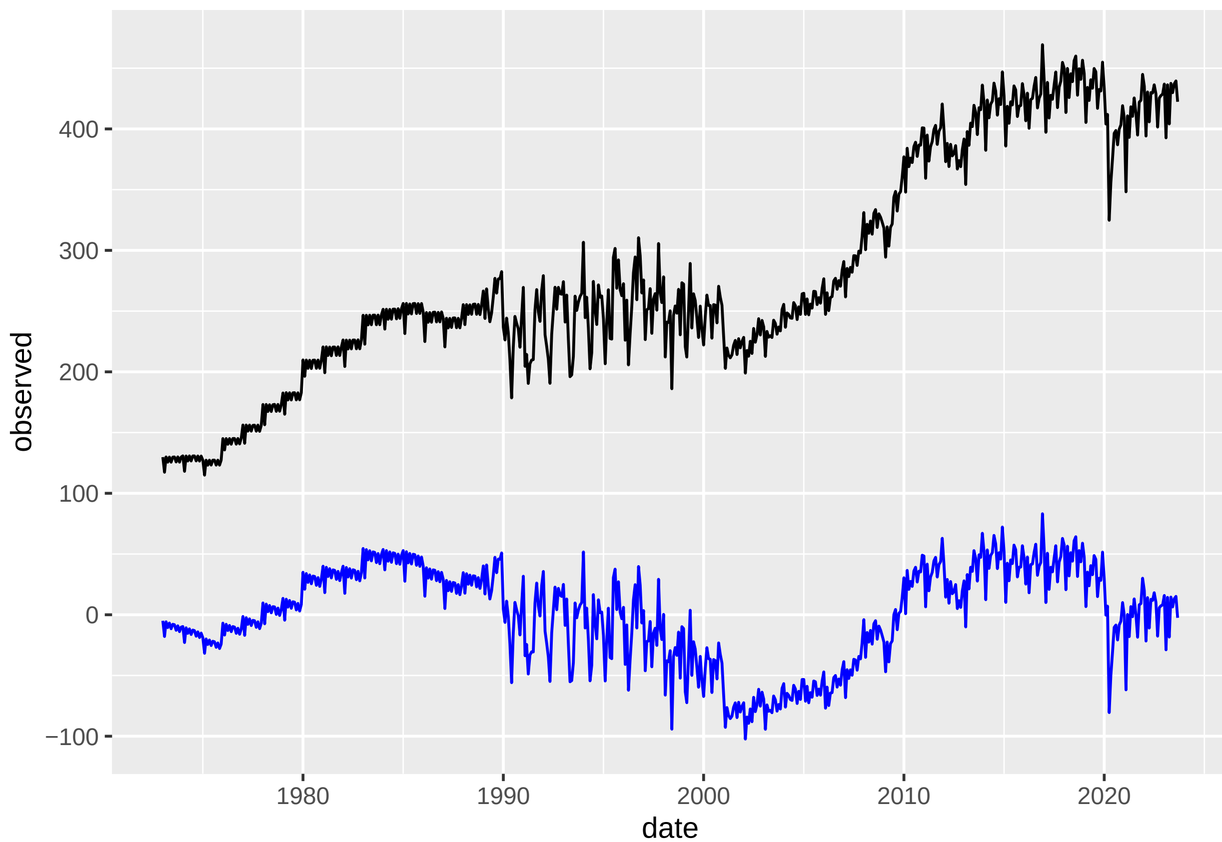
P-values The p-values in this case are small enough, which suggests that both coefficients are significant, indicating a linear trend.

Model fit since both r^2 and f-stats are relatively small, only a small portion of the variability in Hydroelectric Power Consumption is accounted for by the linear trend model.

Q4

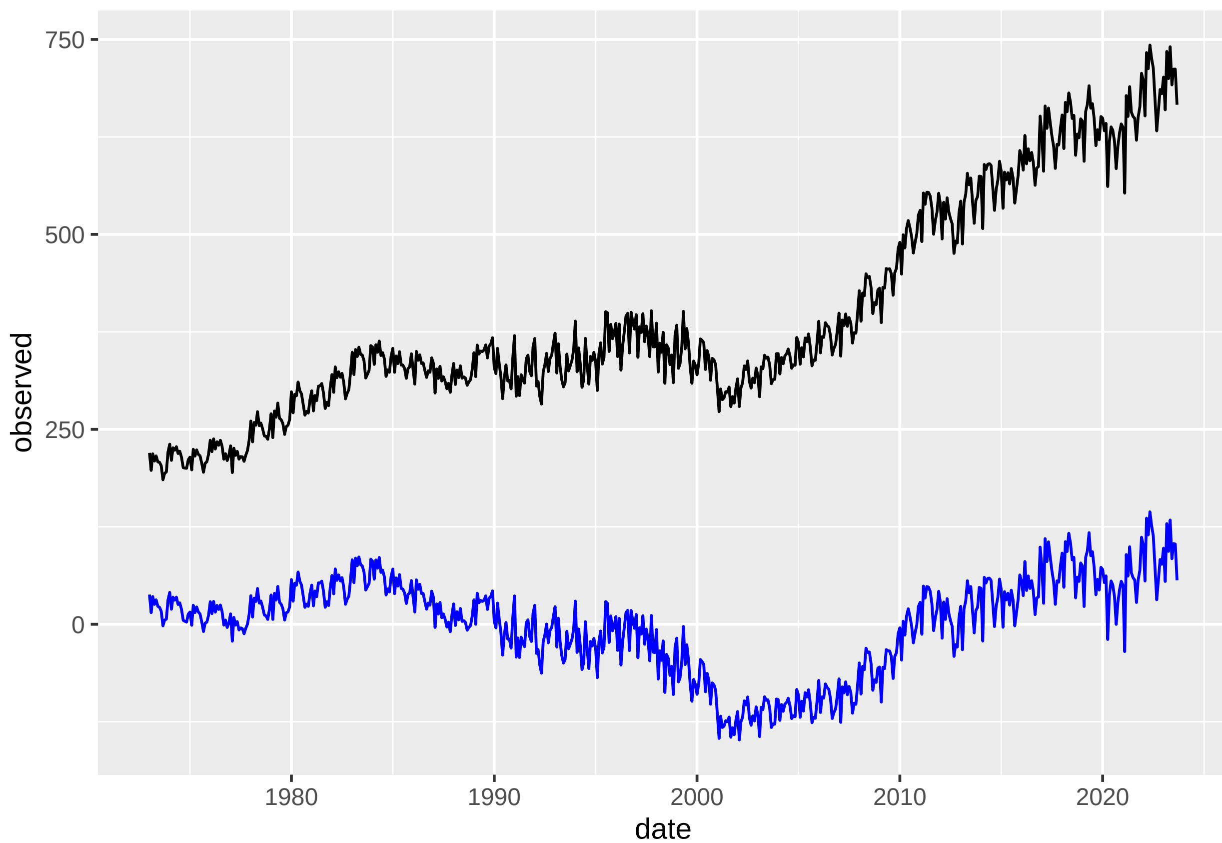
Use the regression coefficients from Q3 to detrend the series. Plot the detrended series and compare with the plots from Q1. What happened? Did anything change?

```
#detrend biomass
bio_detrend <- energy_data[,2] - (beta0_bio + beta1_bio*t)
df_detrend_bio <- data.frame("date" = energy_data$ym_date,
                             "observed" = energy_data[,2],
                             "detrend" = bio_detrend
                             )
ggplot(df_detrend_bio, aes(x=date)) +
  geom_line(aes(y=observed), color = "black") +
  geom_line(aes(y=detrend), color = "blue")
```

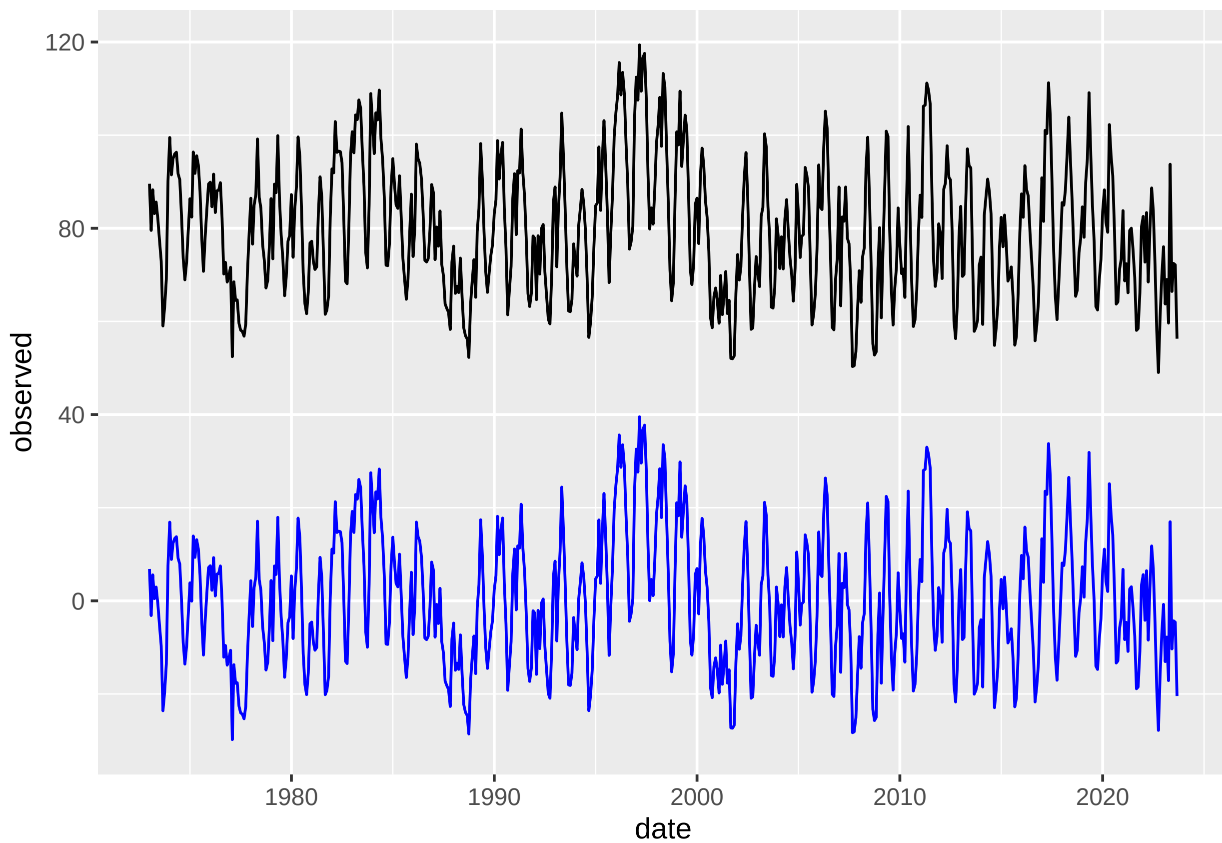



```
#detrend renewable
rnw_detrend <- energy_data[,3] - (beta0_rnw + beta1_rnw*t)
df_detrend_rnw <- data.frame("date" = energy_data$ym_date,
                             "observed" = energy_data[,3],
```

```
                                "detrend" = rnw_detrend
                                )
ggplot(df_detrend_rnw, aes(x=date)) +
  geom_line(aes(y=observed), color = "black") +
  geom_line(aes(y=detrend), color = "blue")
```




```
#detrend hydroelectric
hydro_detrend <- energy_data[,4] - (beta0_hydro + beta1_hydro*t)
df_detrend_hydro <- data.frame("date" = energy_data$ym_date,
                                "observed" = energy_data[,4],
                                "detrend" = hydro_detrend
                                )
ggplot(df_detrend_hydro, aes(x=date)) +
  geom_line(aes(y=observed), color = "black") +
  geom_line(aes(y=detrend), color = "blue")
```



All three plots, after detrending, has become more flat in terms of the long-term trend

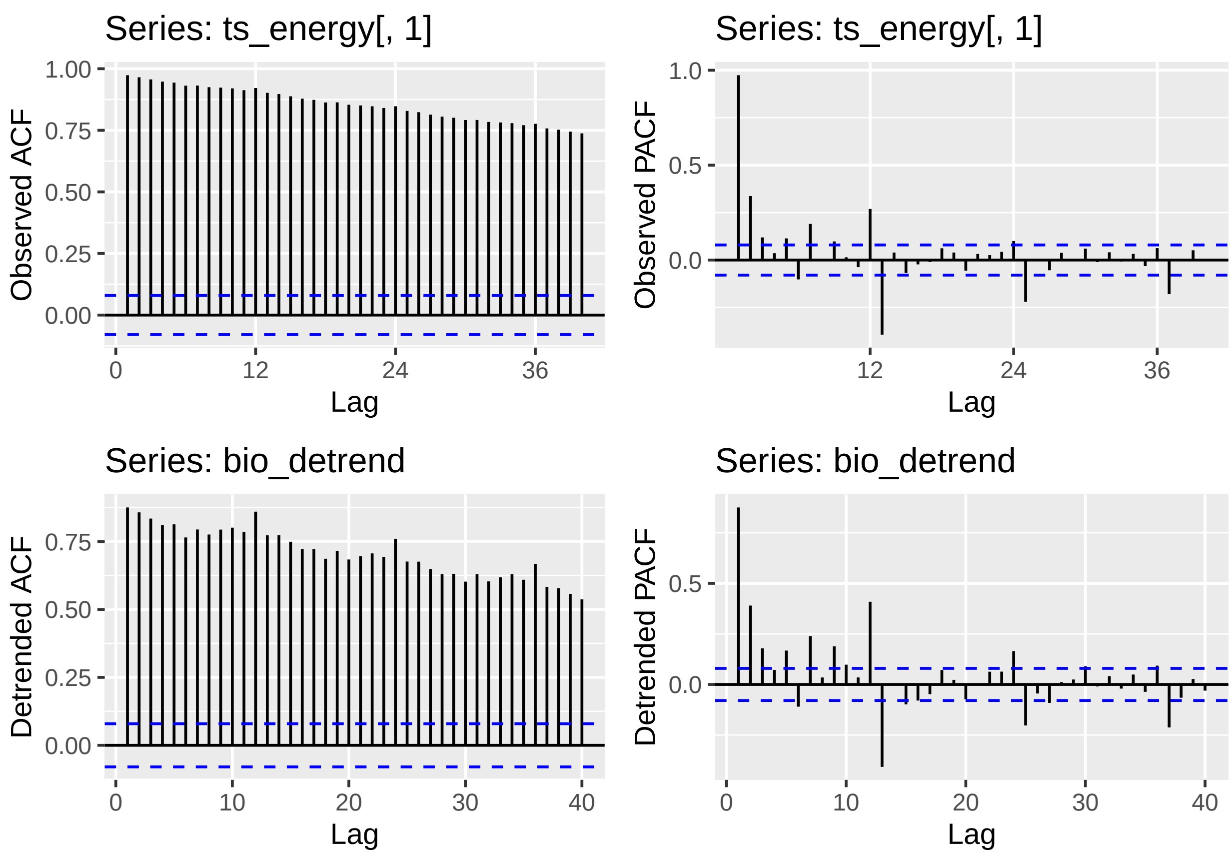
Q5

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. nut mot mandatory. Did the plots change? How?

```
#Biomass detrend plot grid
bio_detrend_plot_grid <- plot_grid(autoplot(biomass_Acf, ylab = "Observed ACF"),
                                   autoplot(biomass_Pacf, ylab = "Observed PACF"),
                                   autoplot(Acf_bio_detrend, ylab = "Detrended ACF"),
                                   autoplot(Pacf_bio_detrend, ylab = "Detrended PACF")
                                   )
```

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

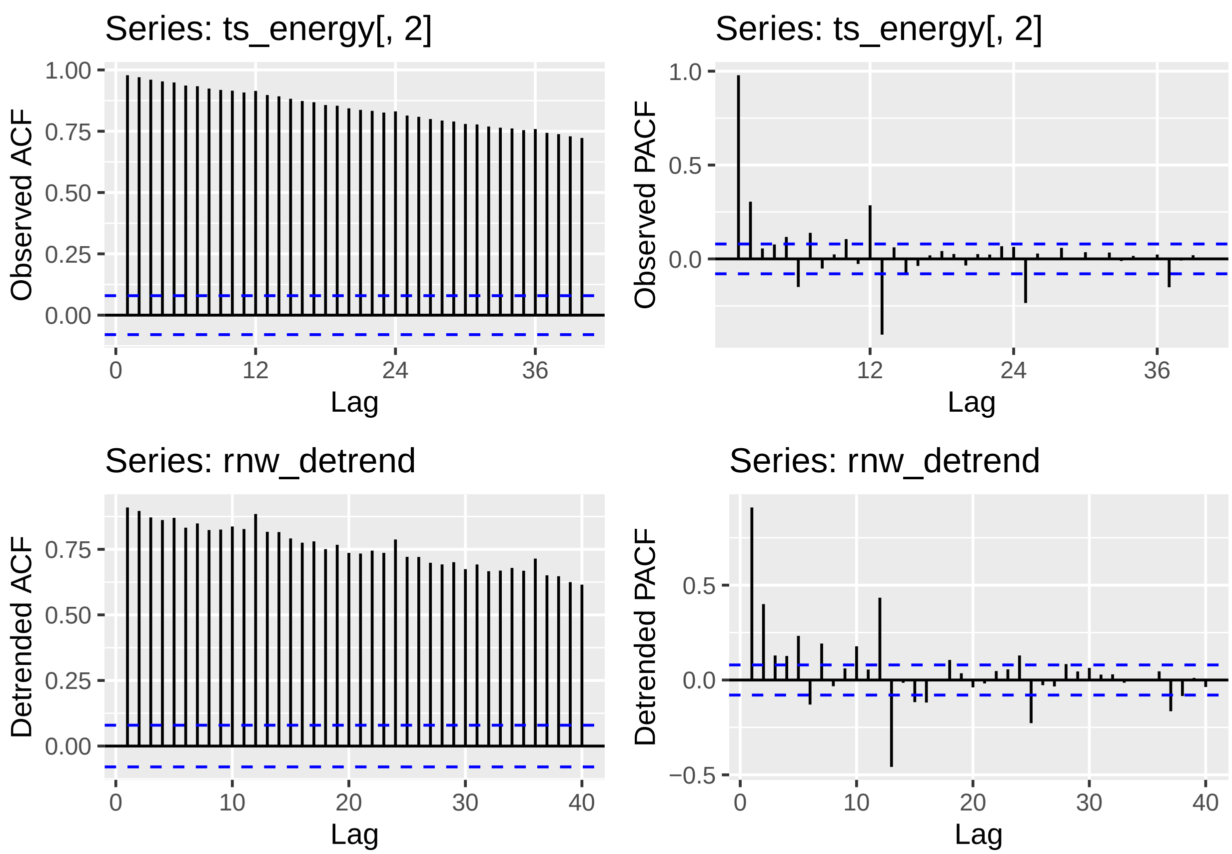
```
bio_detrend_plot_grid
```




```r
#Renewable detrend plot grid
rnw_detrend_plot_grid <- plot_grid(autoplot(rnw_Acf, ylab = "Observed ACF"),
                                   autoplot(rnw_Pacf, ylab = "Observed PACF"),
                                   autoplot(Acf_rnw_detrend, ylab = "Detrended ACF"),
                                   autoplot(Pacf_rnw_detrend, ylab = "Detrended PACF")
                                   )
```

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

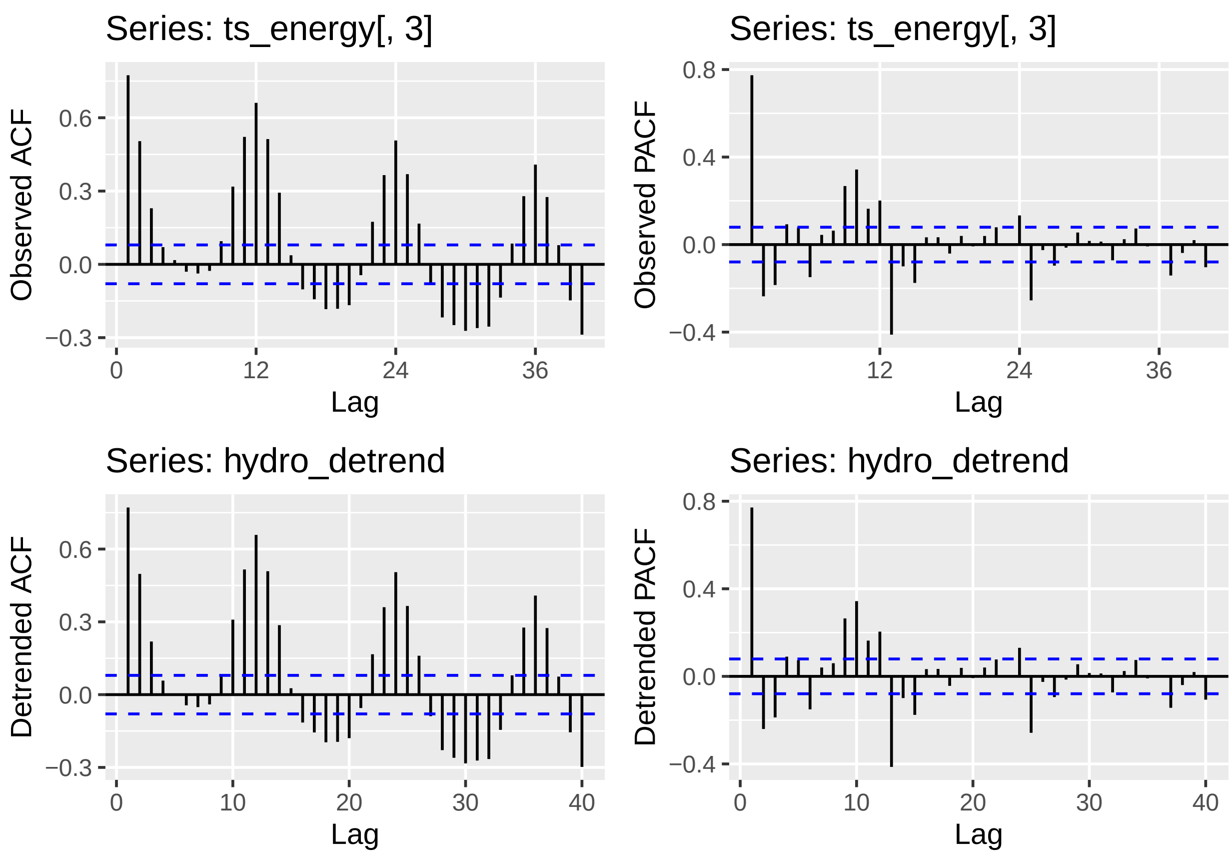
```r
rnw_detrend_plot_grid
```




```r
#Hydroelectric detrend plot grid
hydro_detrend_plot_grid <- plot_grid(autoplot(hydro_Acf, ylab = "Observed ACF"),
                                     autoplot(hydro_Pacf, ylab = "Observed PACF"),
                                     autoplot(Acf_hydro_detrend, ylab = "Detrended ACF"),
                                     autoplot(Pacf_hydro_detrend, ylab = "Detrended PACF")
                                     )
```

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

```r
hydro_detrend_plot_grid
```



The plots of ACF and PACF for total biomass and total renewables has changed a little after detrending, but the change of plots for hydroelectric is not obvious.

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.

For total biomass energy production and total renewable energy production, although there seem to exist regular and consistent patterns occurring at fixed intervals, their acfs exhibit gradually decaying trend instead of having regularly repeated spikes, indicating that there may not exist seasonality. For hydroelectric power consumption, not only does the time series plot have a regular and consistent patterns occuring at fixed intervals, but its acf also has regularly reapted spikes, suggesting that the hydro data has a seasonality.

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
#Create dummies
dummies_rnw <- seasonaldummy(ts_energy[,2])
dummies_hydro <- seasonaldummy(ts_energy[,3])

#Regress on dummies
seas_lm_rnw <- lm(energy_data[,3]~dummies_rnw)
summary(seas_lm_rnw)
```

```
##
## Call:
## lm(formula = energy_data[, 3] ~ dummies_rnw)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -199.19  -86.35  -48.84  113.18  331.58
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     404.526     19.574  20.666   <2e-16 ***
## dummies_rnwJan    2.962     27.546   0.108    0.914
## dummies_rnwFeb  -34.476     27.546  -1.252    0.211
## dummies_rnwMar    3.929     27.546   0.143    0.887
## dummies_rnwApr   -8.695     27.546  -0.316    0.752
## dummies_rnwMay    6.645     27.546   0.241    0.809
## dummies_rnwJun   -4.198     27.546  -0.152    0.879
## dummies_rnwJul    2.460     27.546   0.089    0.929
## dummies_rnwAug   -5.026     27.546  -0.182    0.855
## dummies_rnwSep  -29.119     27.546  -1.057    0.291
## dummies_rnwOct  -20.068     27.682  -0.725    0.469
## dummies_rnwNov  -20.346     27.682  -0.735    0.463
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 138.4 on 597 degrees of freedom
## Multiple R-squared:  0.009296,   Adjusted R-squared:  -0.008958
## F-statistic: 0.5093 on 11 and 597 DF,  p-value: 0.8976
```

```r
seas_lm_hydro <- lm(energy_data[,4]~dummies_hydro)
summary(seas_lm_hydro)
```

```
##
## Call:
## lm(formula = energy_data[, 4] ~ dummies_hydro)
##
## 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 ***
## dummies_hydroJan    4.807      2.069   2.323  0.02050 *
```

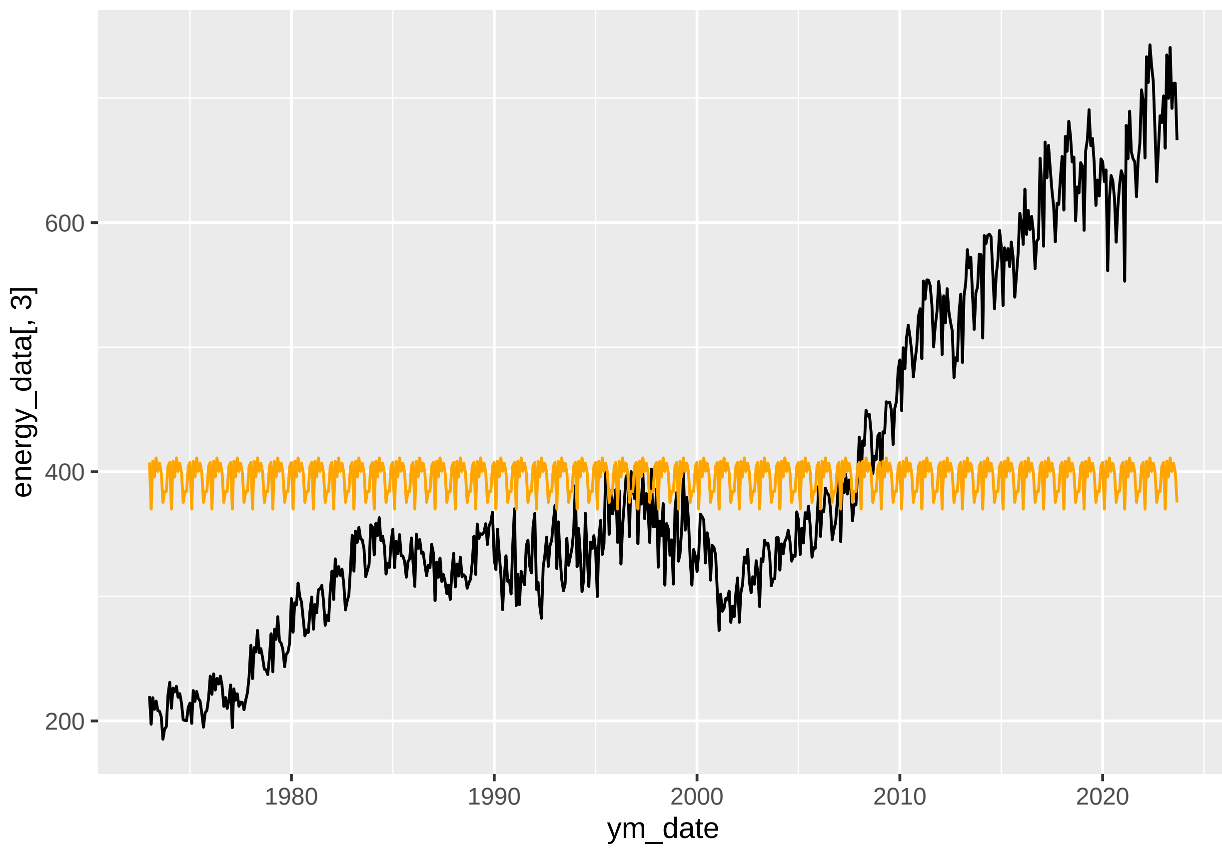
```
## dummies_hydroFeb   -2.725      2.069  -1.317  0.18831
## dummies_hydroMar    6.825      2.069   3.298  0.00103 **
## dummies_hydroApr    5.319      2.069   2.571  0.01039 *
## dummies_hydroMay   13.922      2.069   6.729 4.02e-11 ***
## dummies_hydroJun   10.650      2.069   5.147 3.60e-07 ***
## dummies_hydroJul    3.912      2.069   1.891  0.05914 .
## dummies_hydroAug   -5.677      2.069  -2.744  0.00626 **
## dummies_hydroSep  -16.797      2.069  -8.118 2.72e-15 ***
## dummies_hydroOct  -16.468      2.079  -7.920 1.17e-14 ***
## dummies_hydroNov  -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
```

Based on the summaries from both renewables and hydroelectric, although the dummies in renewables seem to be insignificant, many dummies in hydroelectric appear to be significant, indicating a seasonality in the hydroelectric power consumption time series. The result from both summaries are same as my observation from Q6.

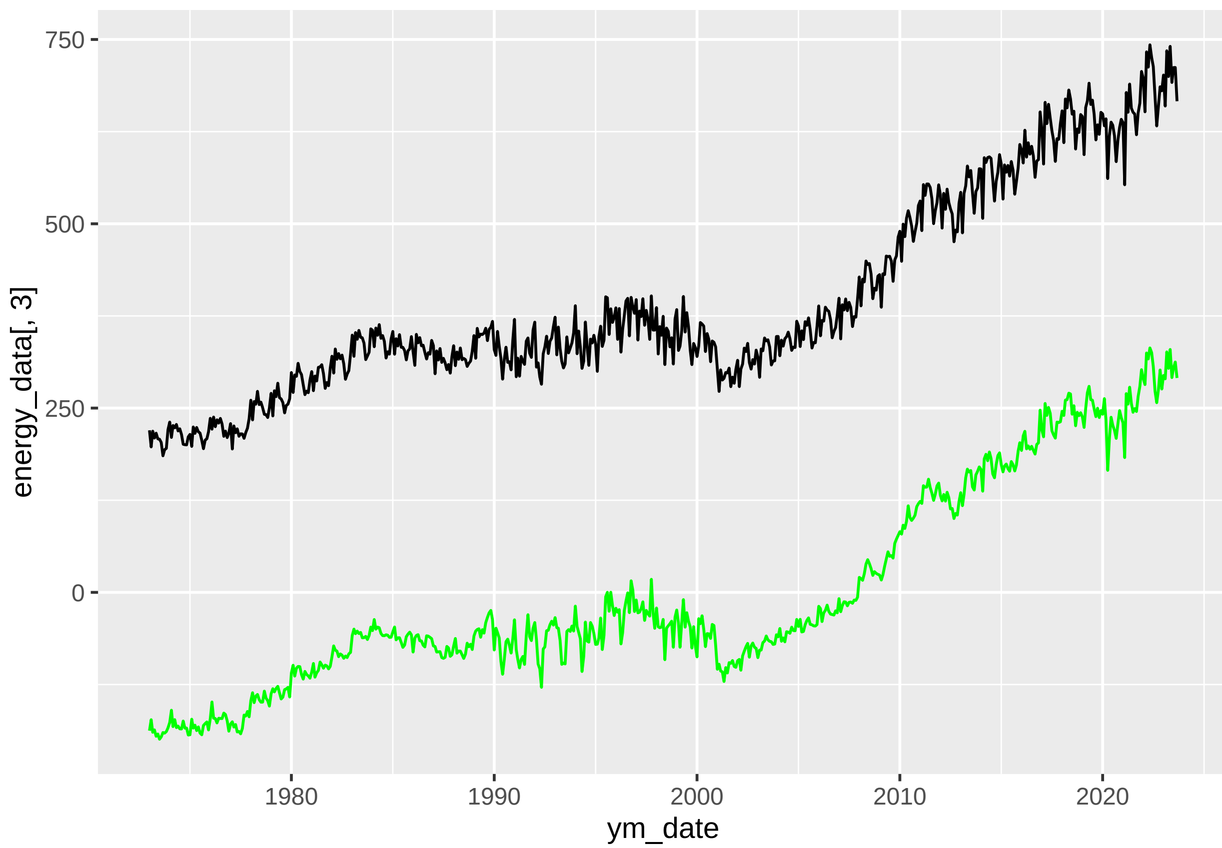
Q8

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?

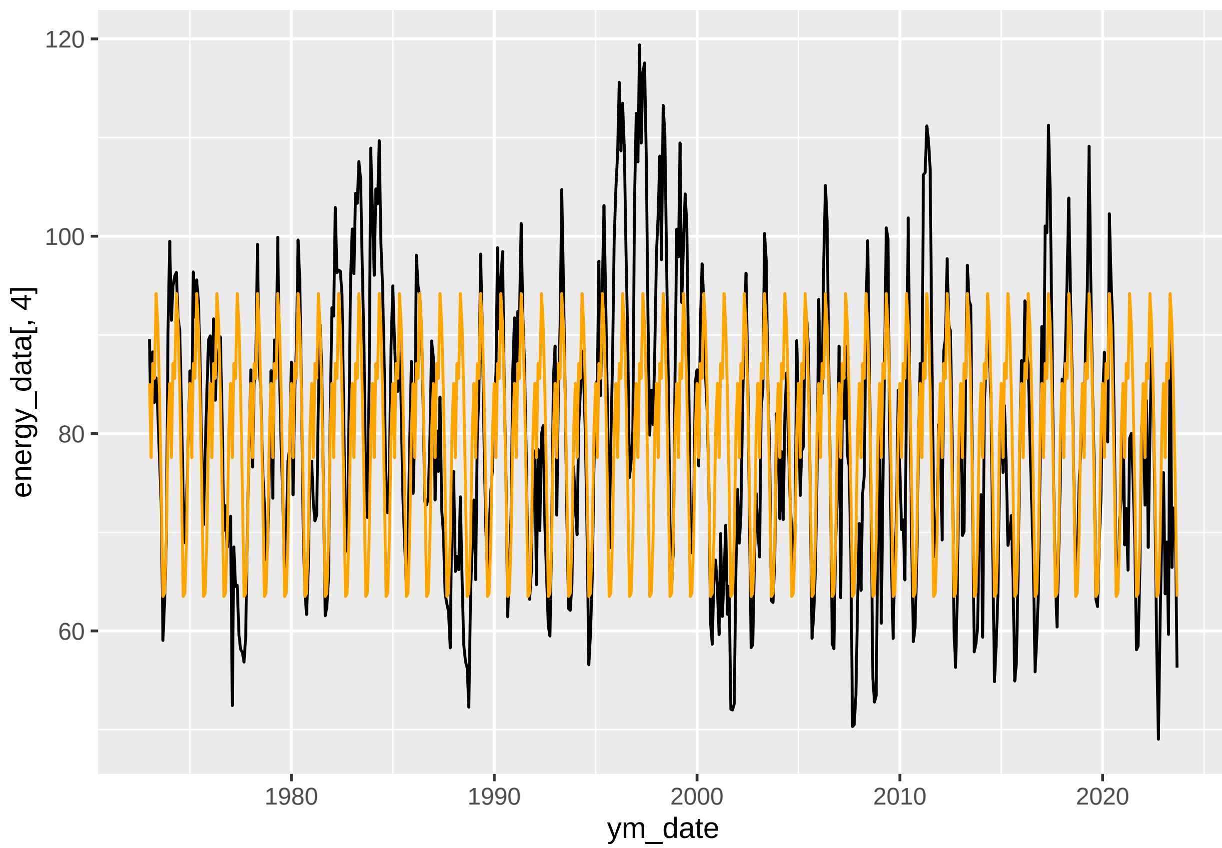
```r
#Deseason process of rnw
#Coeff.
beta0_rnw <- seas_lm_rnw$coefficients[1]
beta1_rnw <- seas_lm_rnw$coefficients[2:12]

#Seasonal comp
seas_comp_rnw <- array(0,n)
for(i in 1:n){
  seas_comp_rnw[i] <- beta0_rnw + beta1_rnw %*% dummies_rnw[i,]
}

ggplot(energy_data, aes(x=ym_date)) +
  geom_line(aes(y=energy_data[,3]), col = "black") +
  geom_line((aes(y=seas_comp_rnw)), col = "orange")
```



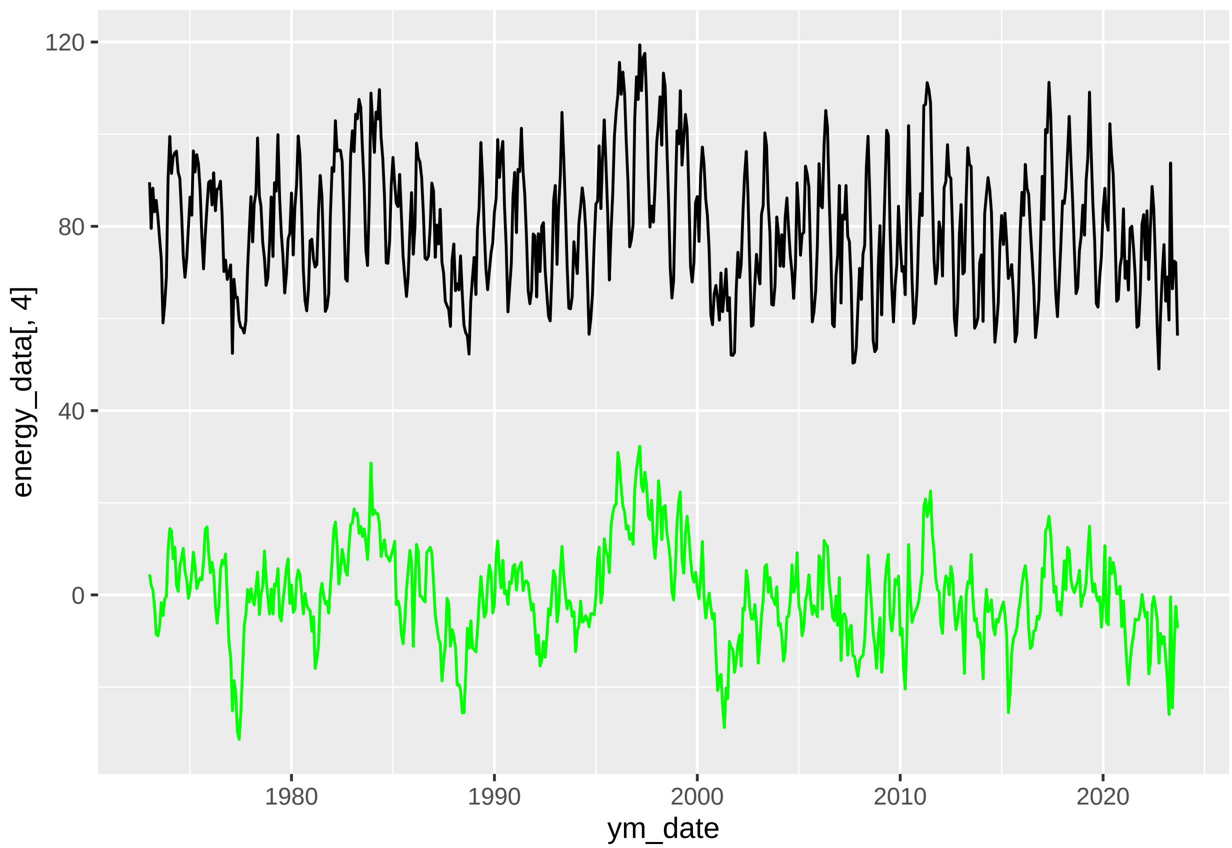

```r
#Detrend
y_deseason_rnw <- energy_data[,3] - seas_comp_rnw

ggplot(energy_data, aes(x = ym_date)) +
  geom_line(aes(y=energy_data[,3]),col="black")+
  geom_line(aes(y=y_deseason_rnw),col="green")
```




```r
#Detrend process of hydro
#Coeff. of hydro
beta0_hydro <- seas_lm_hydro$coefficients[1]
beta1_hydro <- seas_lm_hydro$coefficients[2:12]

#Seasonal comp
seas_comp_hydro <- array(0,n)
for(i in 1:n){
  seas_comp_hydro[i] <- beta0_hydro + beta1_hydro %*% dummies_hydro[i,]
}

ggplot(energy_data, aes(x=ym_date)) +
  geom_line(aes(y=energy_data[,4]), col = "black") +
  geom_line((aes(y=seas_comp_hydro)), col = "orange")
```




```r
#Detrend
y_deseason_hydro <- energy_data[,4] - seas_comp_hydro

ggplot(energy_data, aes(x = ym_date)) +
  geom_line(aes(y=energy_data[,4]),col="black")+
  geom_line(aes(y=y_deseason_hydro),col="green")
```



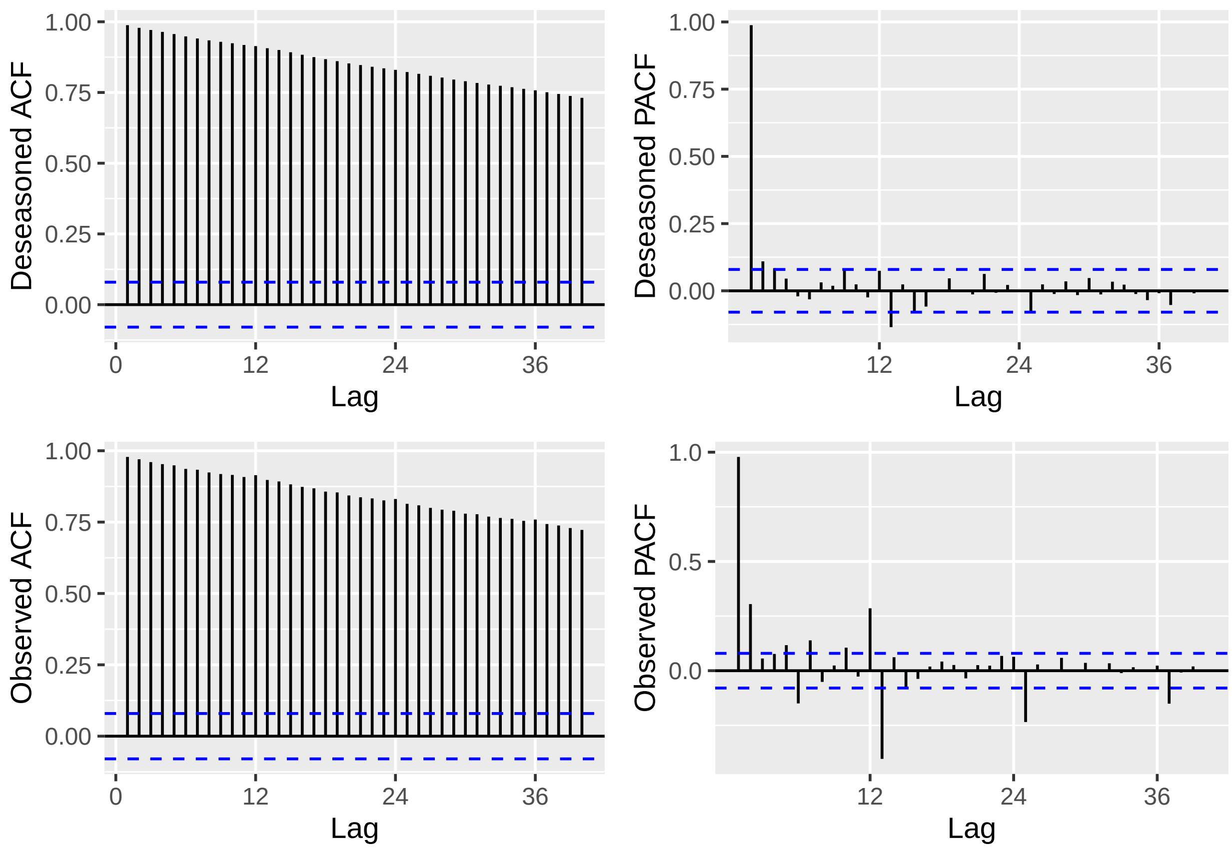
Compared to Q1, while the change in total renewables is not obvious, we can notice some changes in the deseasoned hydroelectric power consumption.

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. nut mot mandatory. Did the plots change? How?

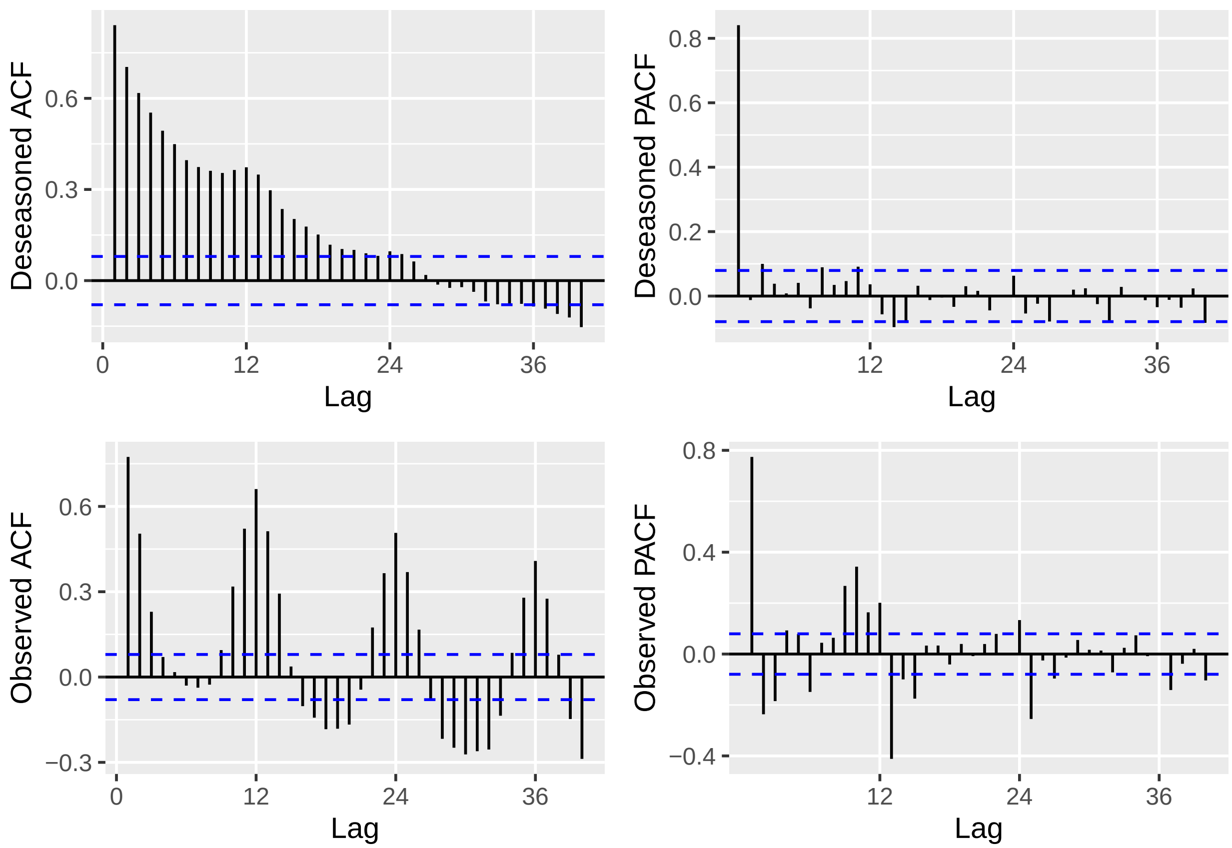
```
#Plot grid for deseason rnw
#Time series data frame of deseasoned datasets
ts_y_deseason_rnw <- ts(y_deseason_rnw, start = c(1973, 1), frequency = 12)

ts_y_deseason_hydro <- ts(y_deseason_hydro, start = c(1973, 1), frequency = 12)

#Deseasoned plots comparison
plot_grid(
  autoplot(Acf(ts_y_deseason_rnw, lag.max = 40, plot = FALSE), ylab = "Deseasoned ACF", main = NULL),
  autoplot(Pacf(ts_y_deseason_rnw, lag.max = 40, plot = FALSE), ylab = "Deseasoned PACF", main = NULL),
  autoplot(Acf(ts_energy[,2], lag.max = 40, plot = FALSE), ylab = "Observed ACF", main = NULL),
  autoplot(Pacf(ts_energy[,2], lag.max = 40, plot = FALSE), ylab = "Observed PACF", main = NULL)
)
```

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




```
plot_grid(
  autoplot(Acf(ts_y_deseason_hydro, lag.max = 40, plot = FALSE),  ylab = "Deseasoned ACF", main = NULL),
  autoplot(Pacf(ts_y_deseason_hydro, lag.max = 40, plot = FALSE), ylab = "Deseasoned PACF", main = NULL),
  autoplot(Acf(ts_energy[,3], lag.max = 40, plot = FALSE), ylab = "Observed ACF", main = NULL),
  autoplot(Pacf(ts_energy[,3], lag.max = 40, plot = FALSE), ylab = "Observed PACF", main = NULL)
)
```

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



While the ACF and PACF of total renewables didn't change much, we can observe that the plots of hydroelectric have changed significantly. Instead of displaying a seasonal cycle in the ACF, the ACF of deseasoned hydro displays a decaying trend. For the PACF, instead of having spikes at every 12 lags, now we can only observe significant spikes for the first two lags for the deseasoned PACF.