Assignment 6: Time Series Analysis

Felipe Raby Amadori

OVERVIEW

This exercise accompanies the lessons in Hydrologic Data Analysis on time series analysis

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single pdf file.
5. After Knitting, submit the completed exercise (pdf file) to the dropbox in Sakai. Add your last name into the file name (e.g., "A06_Salk.html") prior to submission.

The completed exercise is due on 11 October 2019 at 9:00 am.

Setup

1. Verify your working directory is set to the R project file,
2. Load the tidyverse, lubridate, trend, and dataRetrieval packages.
3. Set your ggplot theme (can be theme_classic or something else)
4. Load the ClearCreekDischarge.Monthly.csv file from the processed data folder. Call this data frame ClearCreekDischarge.Monthly.

```
knitr::opts_chunk$set(message = FALSE, warning = FALSE)

#Verify your working directory is set to the R project file
getwd()
```

```
## [1] "C:/Users/Felipe/OneDrive - Duke University/1. DUKE/Ramos 3 Semestre/Hydrologic_Data_Analysis"
```

```
#Load the tidyverse, lubridate, trend, and dataRetrieval packages
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------------

## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(trend)
library(dataRetrieval)
library(lubridate)
```

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

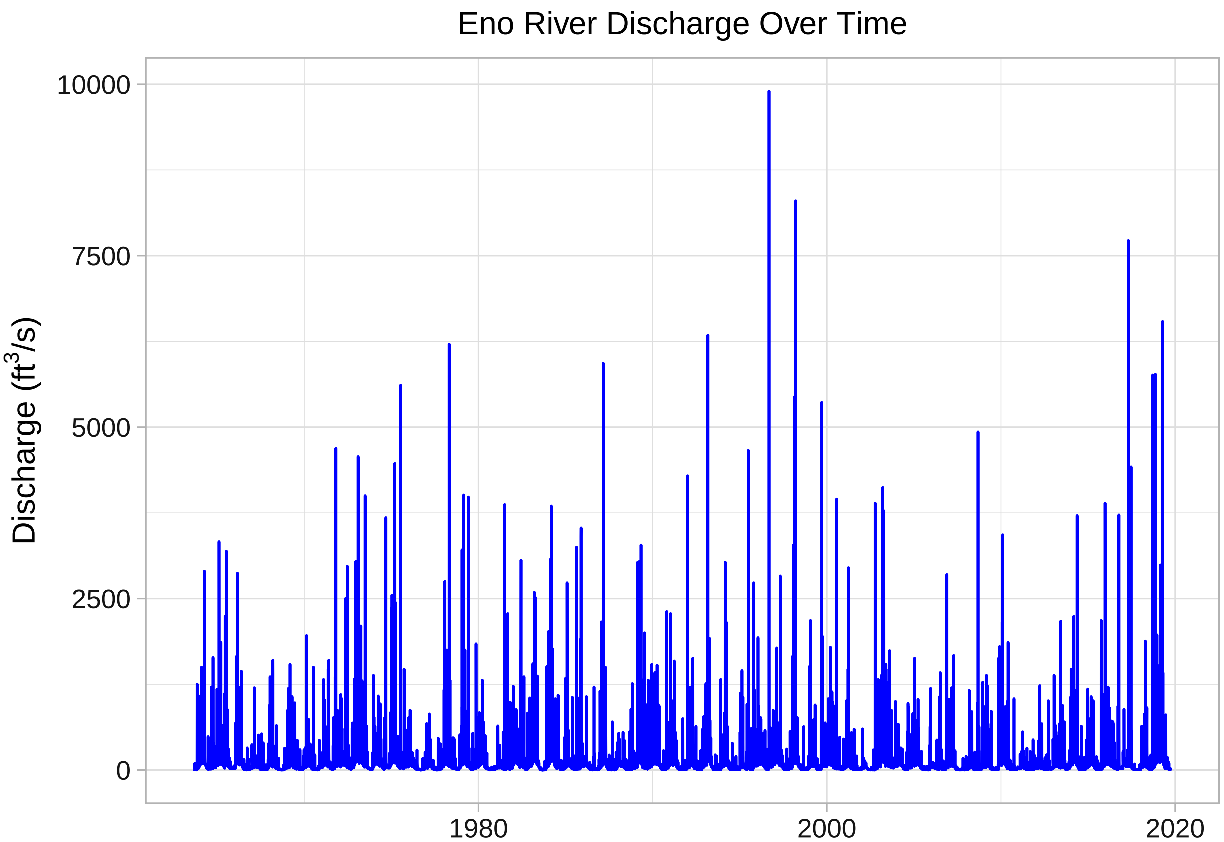
```
## The following object is masked from 'package:base':
##
##     date
```

```r
library(ggplot2)

#Set your ggplot theme
felipe_theme <- theme_light(base_size = 12) +
  theme(axis.text = element_text(color = "grey8"),
        legend.position = "right", plot.title = element_text(hjust = 0.5))
theme_set(felipe_theme)


#Load the ClearCreekDischarge.Monthly.csv file from the processed data folder.
ClearCreekDischarge.Monthly <- read.csv("./Data/Processed/ClearCreekDischarge.Monthly.csv")
```

Time Series Decomposition

5. Create a new data frame that includes daily mean discharge at the Eno River for all available dates (`siteNumbers = "02085070"`). Rename the columns accordingly.
6. Plot discharge over time with geom_line. Make sure axis labels are formatted appropriately.
7. Create a time series of discharge
8. Decompose the time series using the `stl` function.
9. Visualize the decomposed time series.

```r
# Create a new data frame that includes daily mean discharge at the Eno River for all
# available dates (`siteNumbers = "02085070"`). Rename the columns accordingly.
EnoRiverDischarge <- readNWISdv(siteNumbers = "02085070",
                     parameterCd = "00060", # discharge (ft3/s)
                     startDate = "",
                     endDate = "")
names(EnoRiverDischarge)[4:5] <- c("Discharge", "Approval.Code")

# Plot discharge over time with geom_line. Make sure axis labels are formatted appropriately.
EnoRiver.plot <- ggplot(EnoRiverDischarge, aes(x = Date)) +
  geom_line(aes(y = Discharge), color = "blue", show.legend = FALSE) +
  labs(x = "", y = expression("Discharge (ft"^3*"/s)"),
       title = "Eno River Discharge Over Time") +
  theme(plot.title = element_text(size = 12),
        axis.title.x = element_blank())

print(EnoRiver.plot)
```




```
# Create a time series of discharge

# First we check if there are gaps in the data.
table(diff(EnoRiverDischarge$Date))
```

```
##
##     1    39
## 20449     1
```

```
#There is 1 data gap of 38 measurements

# Looking for the dates where there is no measurement
DateRange <- seq(min(EnoRiverDischarge$Date), max(EnoRiverDischarge$Date), by = 1)
DateRange[!DateRange %in% EnoRiverDischarge$Date]
```

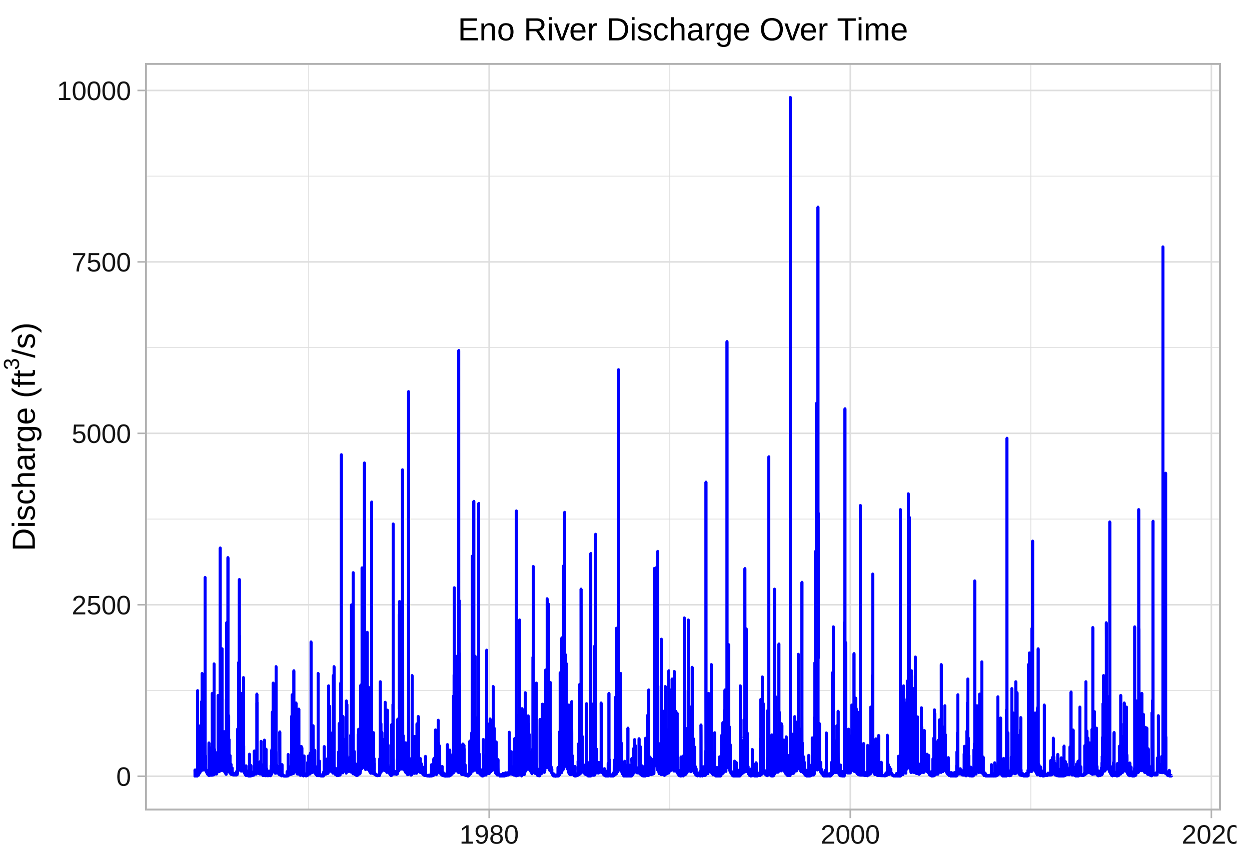
```
##  [1] "2017-10-21" "2017-10-22" "2017-10-23" "2017-10-24" "2017-10-25"
##  [6] "2017-10-26" "2017-10-27" "2017-10-28" "2017-10-29" "2017-10-30"
## [11] "2017-10-31" "2017-11-01" "2017-11-02" "2017-11-03" "2017-11-04"
## [16] "2017-11-05" "2017-11-06" "2017-11-07" "2017-11-08" "2017-11-09"
## [21] "2017-11-10" "2017-11-11" "2017-11-12" "2017-11-13" "2017-11-14"
## [26] "2017-11-15" "2017-11-16" "2017-11-17" "2017-11-18" "2017-11-19"
## [31] "2017-11-20" "2017-11-21" "2017-11-22" "2017-11-23" "2017-11-24"
## [36] "2017-11-25" "2017-11-26" "2017-11-27"
```

```
#Data gap between 10/21/2017 & 11/27/2017

#Considering that time series objects must be equispaced, that the gap is to big to interpolate
#using the neighboring values, and that # after 11/27/2017 we only have less than two years of data
```

```r
#of an approx. 56 years dataset, we are only going to consider for the analysis
#the data from 09/01/1963 to 10/20/2017.

#Filtering the data
EnoRiverDischarge2 <- filter(EnoRiverDischarge, Date < "2017-10-21")

#Checking for gaps
table(diff(EnoRiverDischarge2$Date))
```

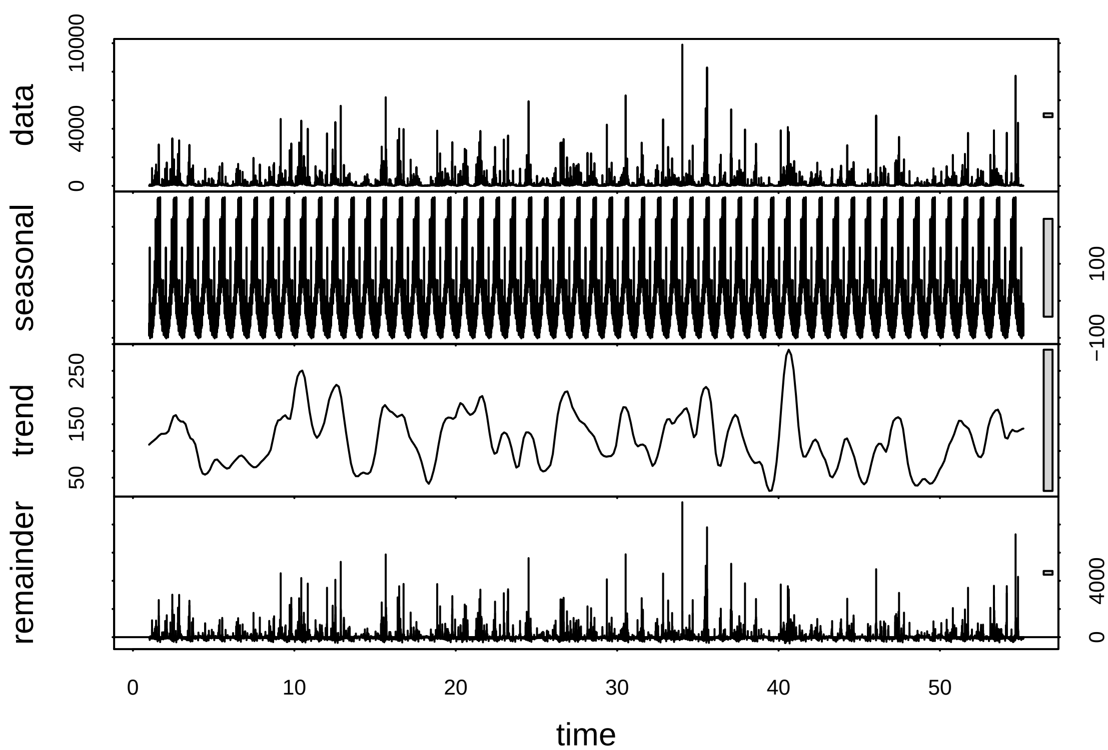
```
##
##     1
## 19773
```

```r
#Plot of the selected data
EnoRiver.plot2 <- ggplot(EnoRiverDischarge2, aes(x = Date)) +
  geom_line(aes(y = Discharge), color = "blue", show.legend = FALSE) +
  labs(x = "", y = expression("Discharge (ft"^3*"/s)"),
       title = "Eno River Discharge Over Time") +
  theme(plot.title = element_text(size = 12),
        axis.title.x = element_blank())

print(EnoRiver.plot2)
```

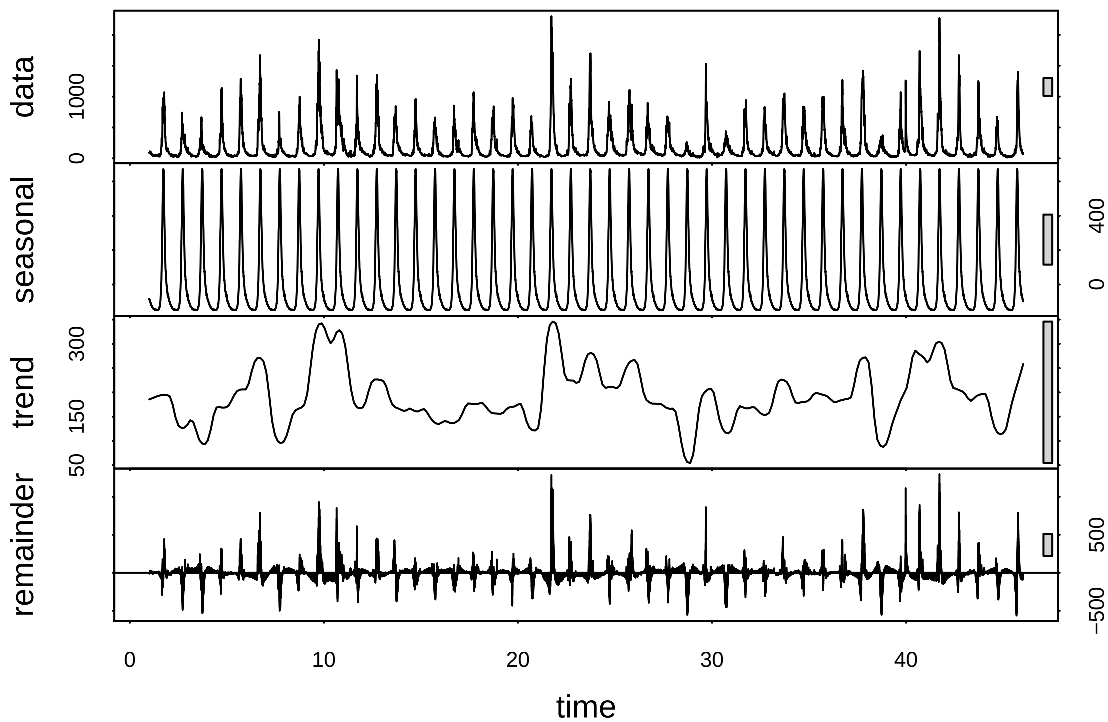



```r
#Time series of discharge
EnoRiver_ts <- ts(EnoRiverDischarge2[[4]], frequency = 365)

#Decompose the time series using the `stl` function.
```

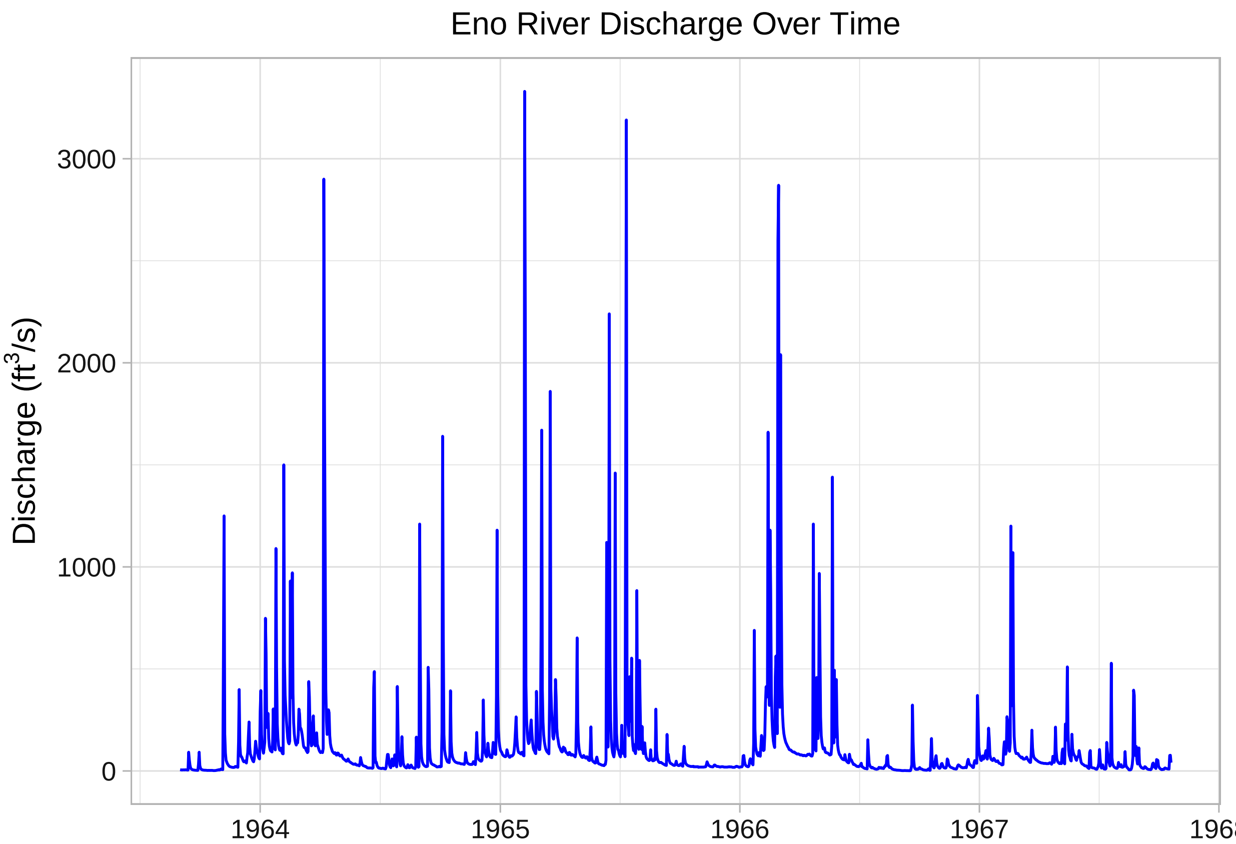
```r
EnoRiver_Decomposed <- stl(EnoRiver_ts, s.window = "periodic")



#From class 11
# Import data
ClearCreekDischarge <- readNWISdv(siteNumbers = "06719505",
                     parameterCd = "00060", # discharge (ft3/s)
                     startDate = "",
                     endDate = "")
names(ClearCreekDischarge)[4:5] <- c("Discharge", "Approval.Code")
ClearCreek_ts <- ts(ClearCreekDischarge[[4]], frequency = 365)
ClearCreek_Decomposed <- stl(ClearCreek_ts, s.window = "periodic")


#Visualize the decomposed Eno River time series.
plot(EnoRiver_Decomposed)
```




```r
# Visualize the decomposed Clear Creek series.
plot(ClearCreek_Decomposed)
```




```
# Zooming to check what is the smaller peak in the seasonal plot
EnoRiverDischarge3 <- filter(EnoRiverDischarge, Date < "1967-10-21")
#Plot of the selected data
EnoRiver.plot3 <- ggplot(EnoRiverDischarge3, aes(x = Date)) +
  geom_line(aes(y = Discharge), color = "blue", show.legend = FALSE) +
  labs(x = "", y = expression("Discharge (ft"^3*"/s)"),
       title = "Eno River Discharge Over Time") +
  theme(plot.title = element_text(size = 12),
        axis.title.x = element_blank())

print(EnoRiver.plot3)
```



10. How do the seasonal and trend components of the decomposition compare to the Clear Creek discharge dataset? Are they similar in magnitude?

#data, seasonal info, trend, remainder #Trend is all over the place. #remainder. All the error that is not predicted by season or trend, #the trends coincide with the high discharge. The trend is always positive. #There are some negative seasonal values.

> Seasonal: There is a clear seasonality in the Eno River and the Clear Creek discharge datasets. Nevertheless, the Eno River data has a double seasonal peak. It has a smaller peak in between the big peaks. It has the bigger seasonal peak in the first months of the year and then it has a smaller seasonal peak in spring/Summer. They are not similar in magnitude probably because the discharge data of the two rivers is very different in magnitude as well.

> Trend: Both datasets do not show any clear trend. Even though the magnitude of the Eno River Discharge data is greater than the Clear Creek Discharge data, the trend plots of both rivers have similar magnitude; moreover, Clear Creek has higher local trends (values over 300 ft/s) than the Eno River (only 1 peak over 250 ft/s).

Trend Analysis

Research question: Has there been a monotonic trend in discharge in Clear Creek over the period of study?

11. Generate a time series of monthly discharge in Clear Creek from the ClearCreekDischarge.Monthly data frame. This time series should include just one column (discharge).
12. Run a Seasonal Mann-Kendall test on the monthly discharge data. Inspect the overall trend and the monthly trends.

```r
# The SMK test requires identically distributed data, condition that this dataset complies with.

ClearCreekDischarge.Monthly_ts <- ts(ClearCreekDischarge.Monthly[[3]],
                                      frequency = 12, start = c(1974, 10))
# 12 because we have monthly data.


ClearCreekDischarge.Monthly_trend <- smk.test(ClearCreekDischarge.Monthly_ts)

# Inspect results
ClearCreekDischarge.Monthly_trend
```

```
##
##  Seasonal Mann-Kendall trend test (Hirsch-Slack test)
##
## data:  ClearCreekDischarge.Monthly_ts
## z = 1.6586, p-value = 0.09719
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##      S   varS
##    590 126102
```

```r
summary(ClearCreekDischarge.Monthly_trend)
```

```
##
##  Seasonal Mann-Kendall trend test (Hirsch-Slack test)
##
## data: ClearCreekDischarge.Monthly_ts
## alternative hypothesis: two.sided
##
## Statistics for individual seasons
##
## H0
##                      S  varS    tau      z Pr(>|z|)
## Season 1:   S = 0   35 10449  0.035  0.333 0.739425
## Season 2:   S = 0    4 10450  0.004  0.029 0.976588
## Season 3:   S = 0  204 10450  0.206  1.986 0.047054 *
## Season 4:   S = 0  230 10450  0.232  2.240 0.025081 *
## Season 5:   S = 0  148 10450  0.149  1.438 0.150434
## Season 6:   S = 0   94 10450  0.095  0.910 0.362951
## Season 7:   S = 0  -54 10450 -0.055 -0.518 0.604135
## Season 8:   S = 0  -99 10449 -0.100 -0.959 0.337703
## Season 9:   S = 0  -90 10450 -0.091 -0.871 0.383958
## Season 10:   S = 0  64 11154  0.062  0.597 0.550828
## Season 11:   S = 0  24 10450  0.024  0.225 0.821984
## Season 12:   S = 0  30 10450  0.030  0.284 0.776650
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

13. Is there an overall monotonic trend in discharge over time? If so, is it positive or negative?

    We fail to reject the null hypothesis of no overall monotonic trend in discharge over time with a 5% level of significance (Hirsch-Slack test, $z = 1.6586$, p-value $= 0.09719 > 0.05$). At a 5% level of significance, the sample data does not provide enough evidence to conclude that there is an overall monotonic trend in discharge over time.

S is positive = 590, so if the case the null hypothesis was rejected the trend would have been positive.

14. Are there any monthly monotonic trends in discharge over time? If so, during which months do they occur and are they positive or negative?

    We reject the null hypothesis of no monotonic trend in discharge over time for the Months of March and April with a 5% level of significance (March: Hirsch-Slack test, z = 1.986, Pr(>|z|) = 0.047054 < 0.05. April: Hirsch-Slack test, z = 2.240, Pr(>|z|) = 0.025081 < 0.05). At a 5% level of significance, the sample data does provide enough evidence to conclude that there is a positive monotonic trend in discharge over time in the Months of March and April..

Reflection

15. What are 2-3 conclusions or summary points about time series you learned through your analysis?

    I learned how time series can be useful for detecting seasonal behaviors and trends in a response variable that has been tracked over time. Also, how to use the ts, and stl functions and the importance of studying the data first for gaps and the frequency the data was recorded. Finally I learned about options to solve gap issues in the data to be able to perform a time series analysis.

16. What data, visualizations, and/or models supported your conclusions from 15?

    Basically the whole lab and classes 11 and 12. Every part of them was super useful.

17. Did hands-on data analysis impact your learning about time series relative to a theory-based lesson? If so, how?

    Yes. Working with the data makes you think about about it and doubts that you didn't have in class emerge. Trying to solve those doubts by yourself is a valuable way of learning.

18. How did the real-world data compare with your expectations from theory?

    To my experience with these type of data (in another country), these was way more complete that the data I am used to. From theory, I believed the real data was ok except for that gap in the year 2017 of the Eno River discharge data.