Assignment 7: Time Series Analysis

Curtis Cha

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics 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., "Fay_A07_TimeSeries.Rmd") prior to submission.

The completed exercise is due on Monday, March 14 at 7:00 pm.

Set up

1. Set up your session:

- Check your working directory
- Load the tidyverse, lubridate, zoo, and trend packages
- Set your ggplot theme

```
#1
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

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

```
library(lubridate)
```

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

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

```
library(zoo)
```

```
##
```

```
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(trend)

mytheme <- theme_classic(base_size = 12) +
  theme(axis.text = element_text(color = "black"), legend.position = "right")
```

2. Import the ten datasets from the Ozone_TimeSeries folder in the Raw data folder. These contain ozone concentrations at Garinger High School in North Carolina from 2010-2019 (the EPA air database only allows downloads for one year at a time). Import these either individually or in bulk and then combine them into a single dataframe named `GaringerOzone` of 3589 observation and 20 variables.

```
#2

setwd("../Data/Raw/Ozone_TimeSeries/")


GaringerOzone <- list.files(pattern="*.csv") %>%
  map_df(~read_csv(.))

dim(GaringerOzone)
```

```
## [1] 3589   20
```

Wrangle

3. Set your date column as a date class.

4. Wrangle your dataset so that it only contains the columns Date, Daily.Max.8.hour.Ozone.Concentration, and DAILY_AQI_VALUE.

5. Notice there are a few days in each year that are missing ozone concentrations. We want to generate a daily dataset, so we will need to fill in any missing days with NA. Create a new data frame that contains a sequence of dates from 2010-01-01 to 2019-12-31 (hint: `as.data.frame(seq())`). Call this new data frame Days. Rename the column name in Days to "Date".

6. Use a `left_join` to combine the data frames. Specify the correct order of data frames within this function so that the final dimensions are 3652 rows and 3 columns. Call your combined data frame GaringerOzone.

```
# 3
GaringerOzone$Date <- mdy(GaringerOzone$Date)

# 4
GaringerOzone_cut <- GaringerOzone %>%
  select("Date", "Daily Max 8-hour Ozone Concentration", "DAILY_AQI_VALUE")

# 5
Days <- as.data.frame(seq(as.Date("2010/01/01"), as.Date("2019/12/31"), by = "day"))
colnames(Days) <- "Date"

# 6
GaringerOzone <- left_join(Days, GaringerOzone_cut)
```

```
## Joining, by = "Date"
```

```
colnames(GaringerOzone) <- c("Date", "Daily_O3_PPM", "Daily_O3_AQI")

dim(GaringerOzone)
```

```
## [1] 3652    3
```

Visualize

7. Create a line plot depicting ozone concentrations over time. In this case, we will plot actual concentrations in ppm, not AQI values. Format your axes accordingly. Add a smoothed line showing any linear trend of your data. Does your plot suggest a trend in ozone concentration over time?

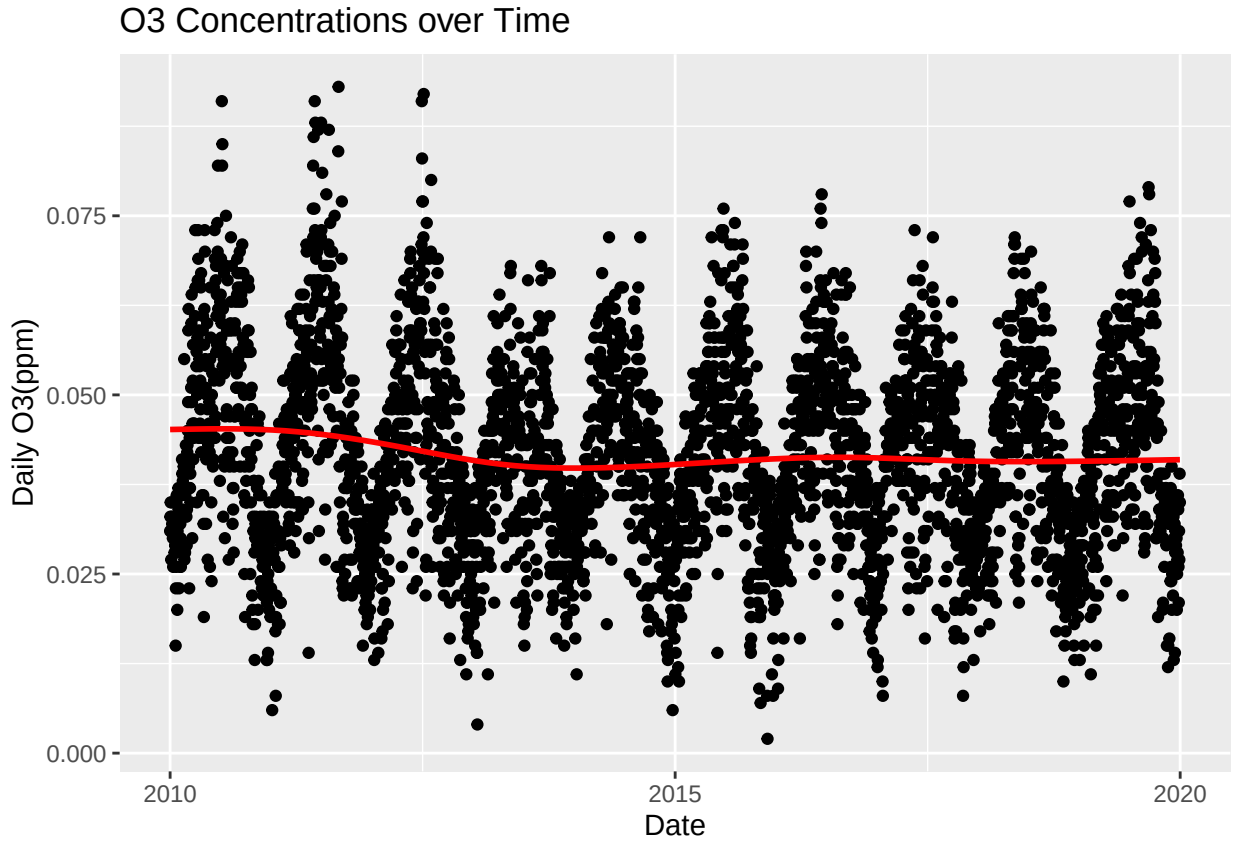
```
#7
library(ggplot2)

ggplot(data = GaringerOzone, aes(x = Date, y = Daily_O3_PPM)) +
  geom_point() + labs(y = "Daily O3(ppm)",
                      title = "O3 Concentrations over Time") +
  geom_smooth(se = F, color = "red")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 63 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 63 rows containing missing values (geom_point).
```



Answer: The red trendline suggests that there is either a weak decrease or no signficant change in

Daily Ozone concentrations over time. While the trendline itself could be viewed as decreasing, it's difficult to determine the actual trend based on visual observation alone.

Time Series Analysis

Study question: Have ozone concentrations changed over the 2010s at this station?

8. Use a linear interpolation to fill in missing daily data for ozone concentration. Why didn't we use a piecewise constant or spline interpolation?

```
#8

GaringerOzone$Daily_O3_PPM <- zoo::na.approx(GaringerOzone$Daily_O3_PPM)
```

Answer: Piece-wise would fill NA's with the nearest member, rather than an in-between value between nearest neighbors. Observing the plot, it's clear that values in sequence are either increasing or decreasing (based on the seasonal pattern of O3 PPM over time). The Spline interpolation uses a quadratic equation for interpolation. It's unlikely that an unknown 03 PPM value shares a quadratic relationship with its two neighbors. For a given day, if O3 PPM the day before is lower than the O3 PPM the day after, it's likely that the O3 PPM of that day is a value in between. A linear interplotion for filling in Daily O3 PPM makes more sense.

9. Create a new data frame called `GaringerOzone.monthly` that contains aggregated data: mean ozone concentrations for each month. In your pipe, you will need to first add columns for year and month to form the groupings. In a separate line of code, create a new Date column with each month-year combination being set as the first day of the month (this is for graphing purposes only)

```
#9

GaringerOzone.monthly <- GaringerOzone %>%
  mutate(Year = year(Date), Month = month(Date)) %>%
  mutate(Date = my(paste(Month,"-",Year))) %>%
  group_by(Date) %>%
  summarise(Monthly_Mean_PPM = mean(Daily_O3_PPM))
```

10. Generate two time series objects. Name the first `GaringerOzone.daily.ts` and base it on the dataframe of daily observations. Name the second `GaringerOzone.monthly.ts` and base it on the monthly average ozone values. Be sure that each specifies the correct start and end dates and the frequency of the time series.

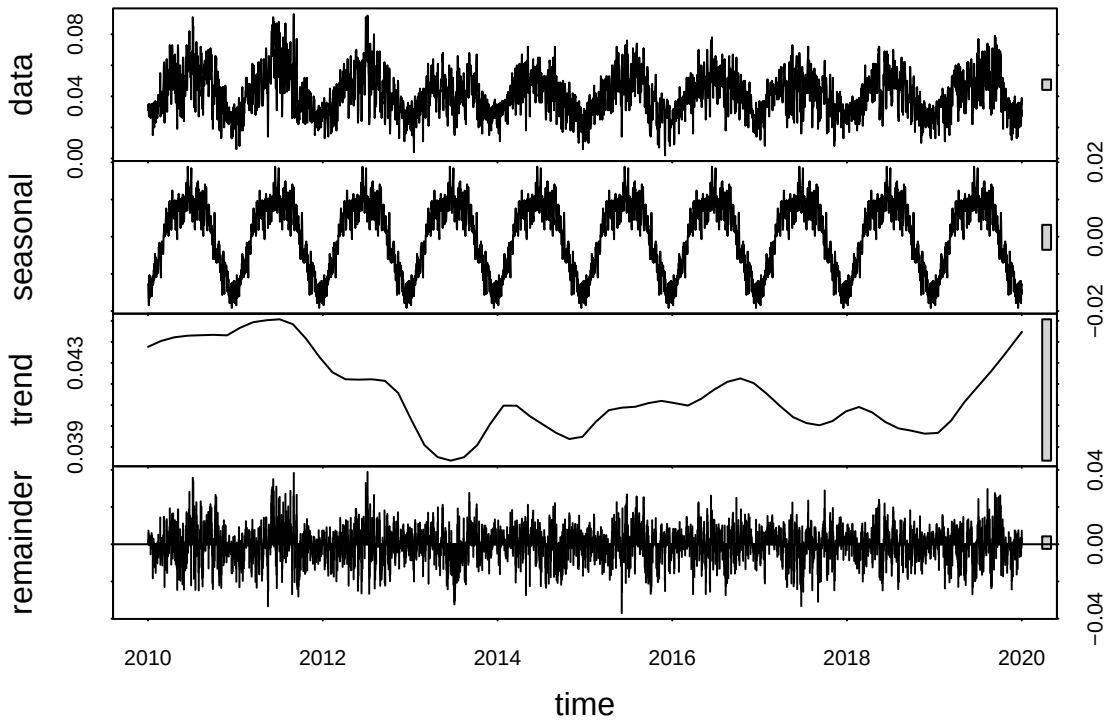
```
#10

GaringerOzone.daily.ts <- ts(data = GaringerOzone$Daily_O3_PPM,
                             start = c(2010,1), frequency = 365)
GaringerOzone.monthly.ts <- ts(data = GaringerOzone.monthly$Monthly_Mean_PPM,
                             start = c(2010,1), frequency = 12)
```

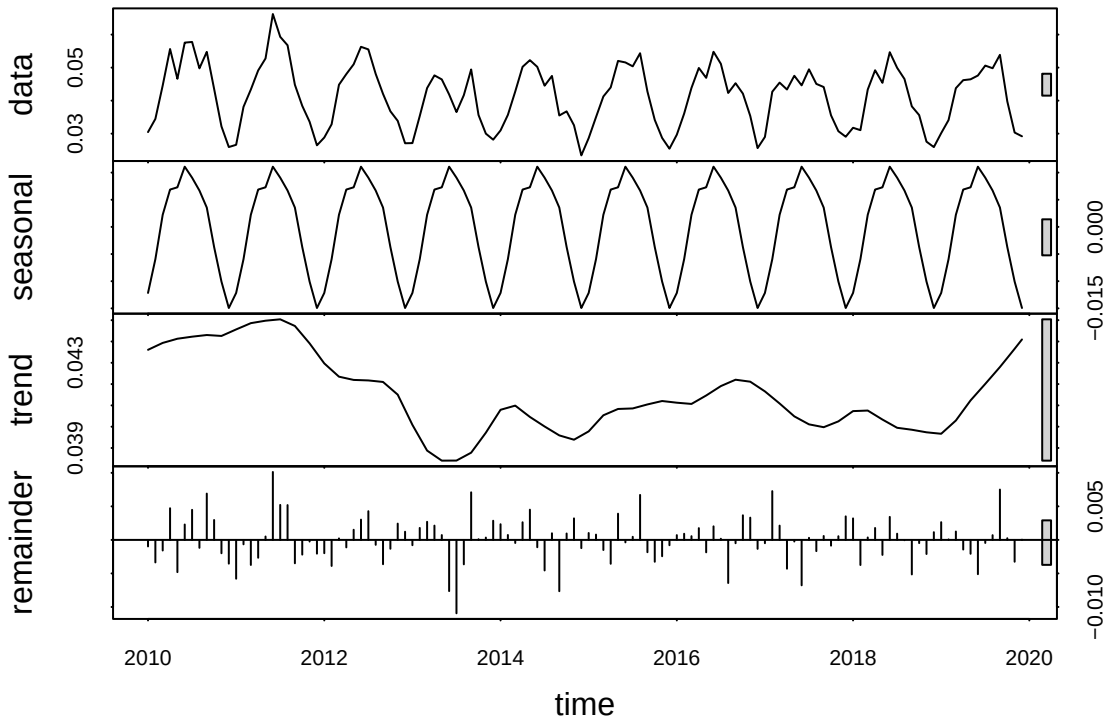
11. Decompose the daily and the monthly time series objects and plot the components using the `plot()` function.

```
#11

GaringerOzone.daily.decomp <- stl(GaringerOzone.daily.ts,
                                  s.window = "periodic")
plot(GaringerOzone.daily.decomp)
```




```
GaringerOzone.monthly.decomp <- stl(GaringerOzone.monthly.ts,
                                    s.window = "periodic")
plot(GaringerOzone.monthly.decomp)
```



12. Run a monotonic trend analysis for the monthly Ozone series. In this case the seasonal Mann-Kendall is most appropriate; why is this?

```
#12

library(Kendall)
summary(Kendall::SeasonalMannKendall(GaringerOzone.monthly.ts))
```

```
## Score =  -77 , Var(Score) = 1499
## denominator =  539.4972
## tau = -0.143, 2-sided pvalue =0.046724
```

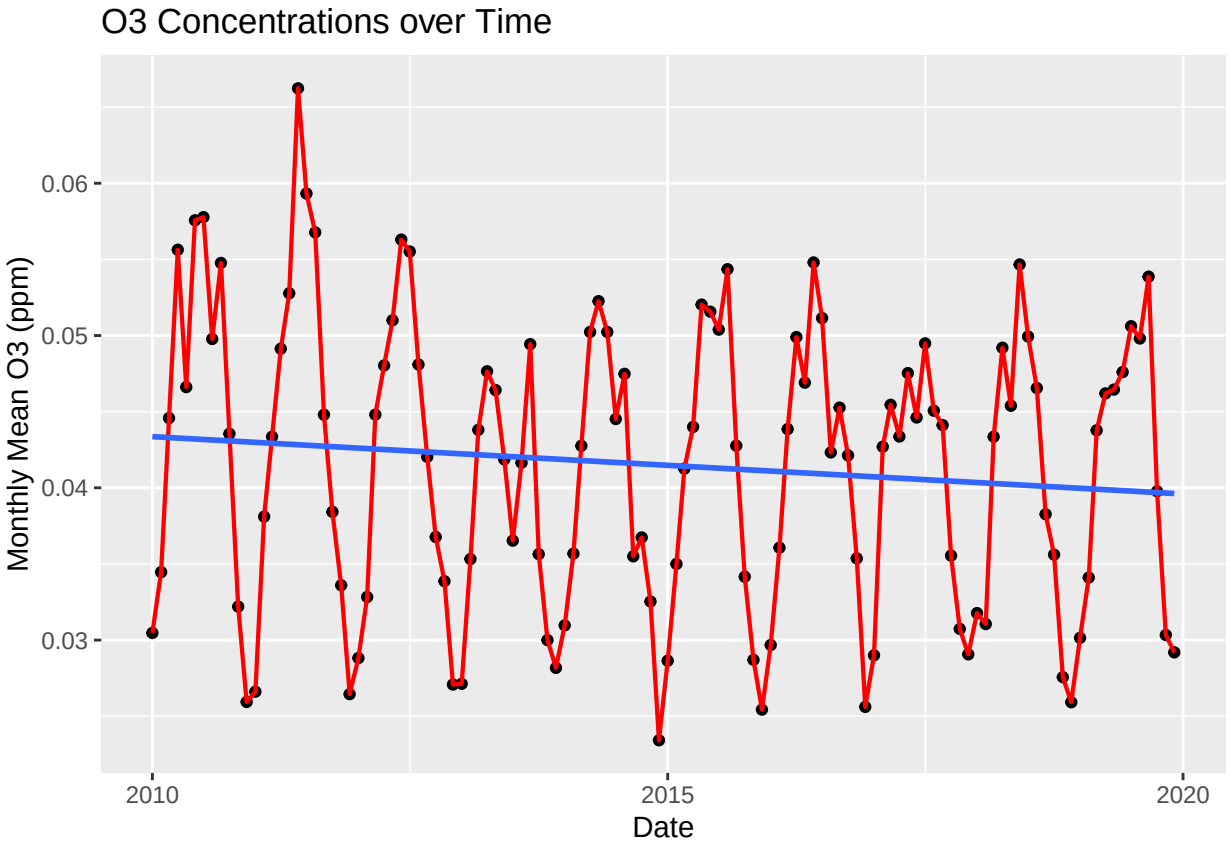
Answer: The Seasonal Mann-Kendall is appropriate because the scatter plot of O3 over time depicts a seasonal pattern. The other four tests (linear reg, Mann-Kendall (non-seasonal), Rho, Aug. Dickey-Fuller) are NOT useful for analyzing seasonality.

13. Create a plot depicting mean monthly ozone concentrations over time, with both a geom_point and a geom_line layer. Edit your axis labels accordingly.

```
# 13

ggplot(data = GaringerOzone.monthly, aes(x = Date, y = Monthly_Mean_PPM)) +
  geom_point() + labs(y = "Monthly Mean O3 (ppm)",
                      title = "O3 Concentrations over Time") +
  geom_line(color = "red", size = 0.75) +
  geom_smooth(method = "lm", se = F)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



14. To accompany your graph, summarize your results in context of the research question. Include output from the statistical test in parentheses at the end of your sentence. Feel free to use multiple sentences in your interpretation.

    Answer: The graph of mean monthly O3 concentrations between 2010 and 2020 demonstrate a weak, negative relationship between O3 (ppm) and time. Visually, there is seasonality of O3 concentrations. As time approaches winter months, there are lower O3 mean values. In contrast, O3 values rise as time approaches summer months. The seasonal MK test also provides statistical support for a negative relationship between O3 and time. Results of seasonal MK test: score = -77, tau = -0.143, 2-sided pvalue = 0.046724. Since the p-value is extremely low, we reject the null hypothesis that time series data is stationary. The negative score and tau indicate that the relationship itself is negative.

15. Subtract the seasonal component from the `GaringerOzone.monthly.ts`. Hint: Look at how we extracted the series components for the EnoDischarge on the lesson Rmd file.

16. Run the Mann Kendall test on the non-seasonal Ozone monthly series. Compare the results with the ones obtained with the Seasonal Mann Kendall on the complete series.

```
#15
GaringerOzone.monthly.Decomposed <- stl(GaringerOzone.monthly.ts,
                                        s.window = "periodic")

GaringerOzone.monthly.noseason <- GaringerOzone.monthly.ts -
  GaringerOzone.monthly.Decomposed$time.series[,1]


#16
```

```
summary(MannKendall(GaringerOzone.monthly.noseason))
```

```
## Score =  -1179 , Var(Score) = 194365.7
## denominator =  7139.5
## tau = -0.165, 2-sided pvalue =0.0075402
```

Answer: After subtracting seasonality from the O3 monthly mean data and running the non-seasonal Mann Kendall test, I observe similar results as that of the seasonal Mann Kendall test. Similar to the seasonal MK test, the non-seasonal test outputs a significantly low p-value and negative tau. We reject the null hypothesis that the non-seasonal O3 data is stationary and observe a negative decreasing trend of O3 over time. However, one difference is that the tau value (and p-value) of the MK test is lower than that of the seasonal MK test. This indicates that the non-seasonal O3 values have a stronger negative relationship with time. (Results of MK: tau = -0.165, 2-sided pvalue = 0.0075402)