Assignment 7: Time Series Analysis

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 Tuesday, March 16 at 11:59 pm.

Set up

1. Set up your session:

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

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.

```
#1
getwd()
```

```
## [1] "/Users/mandyhooks/Environmental_Data_Analytics_2021/Assignments"
```

```
library(tidyverse)
library(zoo)
library(lubridate)
library(trend)
library(dplyr)
library(Kendall)

mytheme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
theme_set(mytheme)

EPAair_O3_GaringerNC2010_raw<-read.csv("/Users/mandyhooks/Environmental_Data_Analytics_2021/Data/Raw/Oz

EPAair_O3_GaringerNC2011_raw <-read.csv("/Users/mandyhooks/Environmental_Data_Analytics_2021/Data/Raw/O

EPAair_O3_GaringerNC2012_raw <-read.csv("/Users/mandyhooks/Environmental_Data_Analytics_2021/Data/Raw/O
```

```
EPAair_O3_GaringerNC2013_raw <-read.csv("/Users/mandyhooks/Environmental_Data_Analytics_2021/Data/Raw/O

EPAair_O3_GaringerNC2014_raw <-read.csv("/Users/mandyhooks/Environmental_Data_Analytics_2021/Data/Raw/O

EPAair_O3_GaringerNC2015_raw <-read.csv("/Users/mandyhooks/Environmental_Data_Analytics_2021/Data/Raw/O

EPAair_O3_GaringerNC2016_raw <-read.csv("/Users/mandyhooks/Environmental_Data_Analytics_2021/Data/Raw/O

EPAair_O3_GaringerNC2017_raw <-read.csv("/Users/mandyhooks/Environmental_Data_Analytics_2021/Data/Raw/O

EPAair_O3_GaringerNC2018_raw <-read.csv("/Users/mandyhooks/Environmental_Data_Analytics_2021/Data/Raw/O

EPAair_O3_GaringerNC2019_raw <-read.csv("/Users/mandyhooks/Environmental_Data_Analytics_2021/Data/Raw/O


#2
EPAair<-rbind(EPAair_O3_GaringerNC2010_raw, EPAair_O3_GaringerNC2011_raw,EPAair_O3_GaringerNC2012_raw, 
```

Wrangle

3. Set your date column as a date class.

```
#3
EPAair$Date<-as.Date(EPAair$Date, format = "%m/%d/%Y")
```

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

```
#4
EPAair<-
  EPAair %>%
  select(Daily.Max.8.hour.Ozone.Concentration, Date, 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".

```
#5
Days <- EPAair %>%
  mutate(Date = as.Date(Date)) %>%
  complete(Date = seq.Date(min(Date), max(Date), by="day"))

view(Days)
```

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.

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

```
## Joining, by = c("Date", "Daily.Max.8.hour.Ozone.Concentration", "DAILY_AQI_VALUE")
```

```
dim(GaringerOzone)
```

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

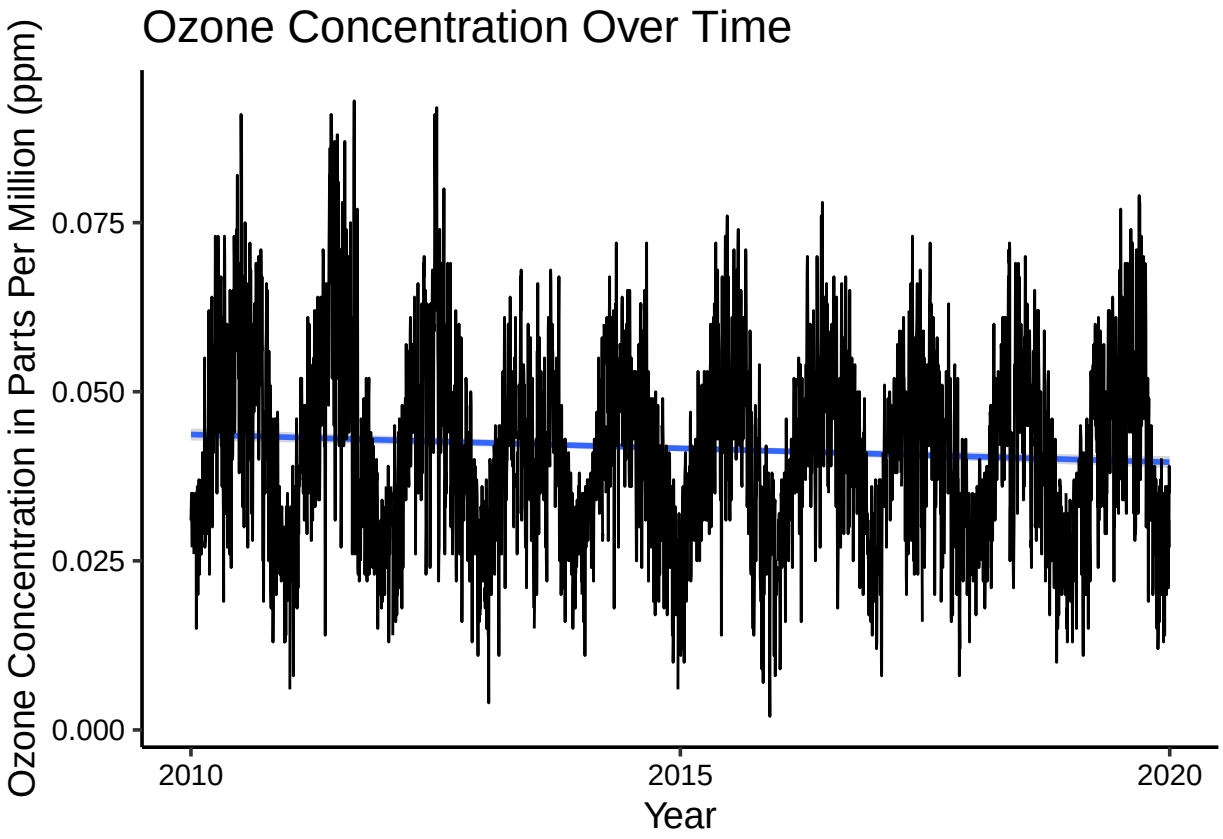
Visualize

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
ozoneplot<-ggplot(GaringerOzone, aes(x=Date, y=Daily.Max.8.hour.Ozone.Concentration))+
  geom_smooth(method=lm)+
  geom_line()+
  labs(x="Year", y= "Ozone Concentration in Parts Per Million (ppm)", title = "Ozone Concentration Over


print(ozoneplot)
```

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

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



Answer: There is a trend, while the ozone seems cyclical, the smoothed line shows an overall decrease in ozone over time.

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 <-
```

```
  GaringerOzone %>%
  mutate( Concentrations_Clean = zoo::na.approx(Daily.Max.8.hour.Ozone.Concentration) )

view(GaringerOzone)
```

Answer: We don't use spline because we are not working with quadratic information, and we want a straight line here rather than a curve of some sort. We do not use piecewise because it would give us a date closest to the data we have, which could give us repeating days. Here, we want linear because we want to fill in the days we are missing and their corresponding data which is most likely, in this case,to be between the data we have now.

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)) %>%
  mutate(Month =month(Date)) %>%
  group_by(Month, Year) %>%
  summarize(mean(Concentrations_Clean)) %>%
  arrange(Year, Month)
```

```
## `summarise()` has grouped output by 'Month'. You can override using the `.groups` argument.
```

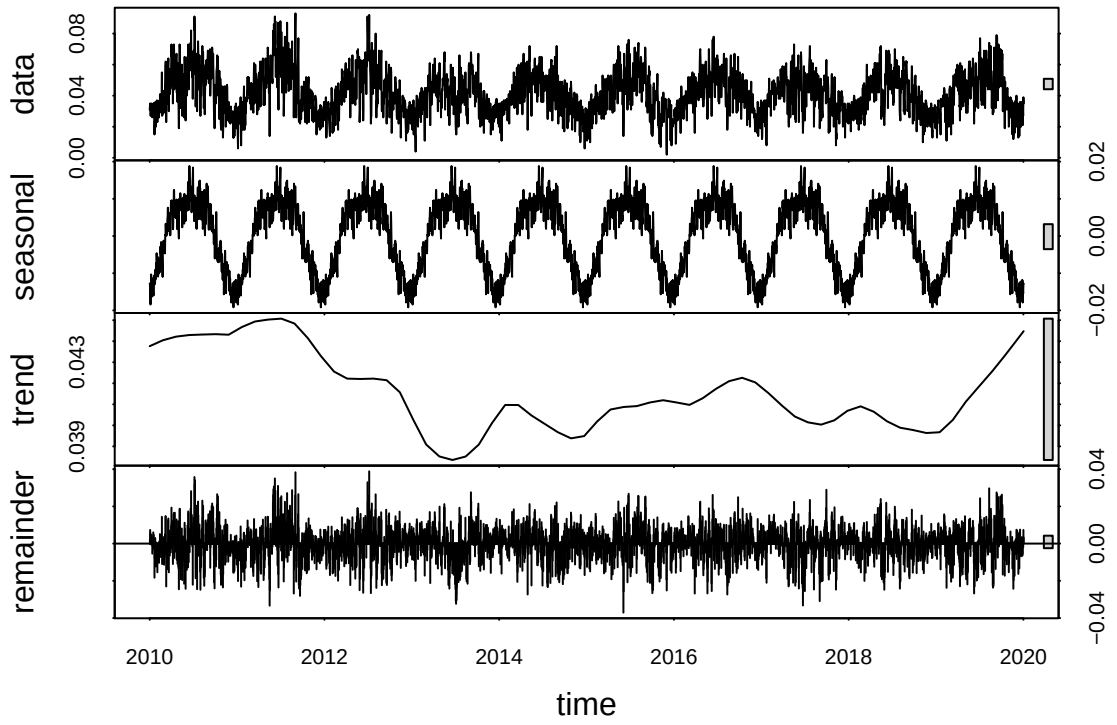
```
GaringerOzone.monthly<-GaringerOzone.monthly %>%
  mutate(Date = my(paste0(Month,"-",Year)))

view(GaringerOzone.monthly)
```

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.

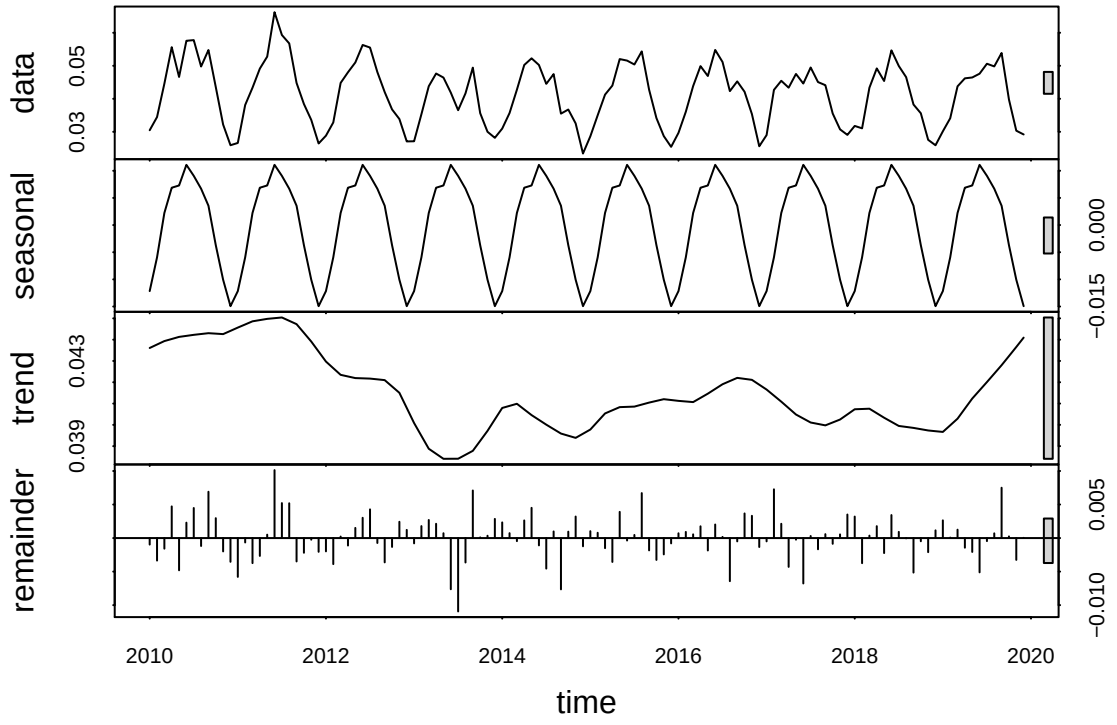
```
#10
GaringerOzone.daily.ts<-ts(GaringerOzone$Concentrations_Clean, start = c(2010,1), frequency = 365)
GaringerOzone.monthly.ts<- ts(GaringerOzone.monthly$`mean(Concentrations_Clean)`, start = c(2010,1), fr

view(GaringerOzone.daily.ts)
```

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

```
#11
GaringerOzone.daily.Decomposed <- stl(GaringerOzone.daily.ts, s.window = "periodic")
GaringerOzone.monthly.Decomposed <- stl(GaringerOzone.monthly.ts, s.window = "periodic")

plot(GaringerOzone.daily.Decomposed)
```




```
plot(GaringerOzone.monthly.Decomposed)
```



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

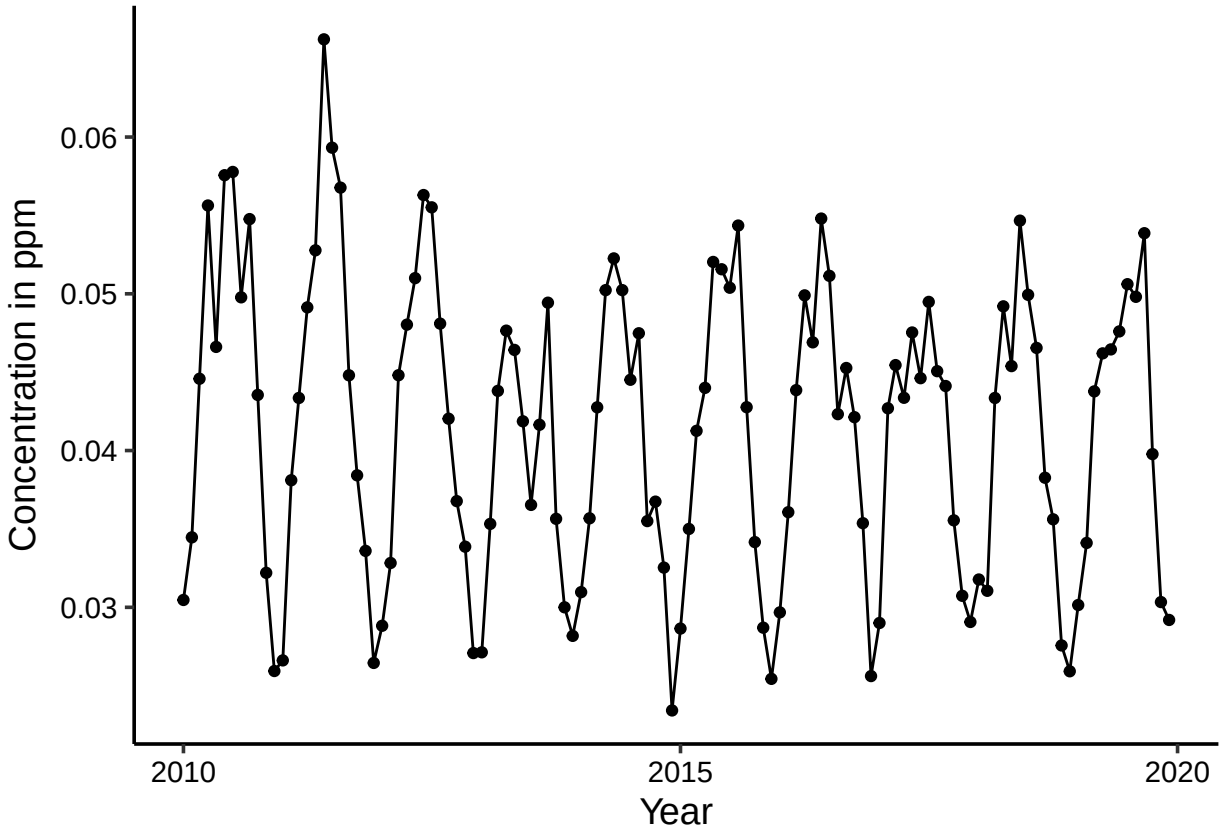
```
#12
Ozone_Trend <- Kendall::SeasonalMannKendall(GaringerOzone.monthly.ts)

summary(Ozone_Trend)
```

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

Answer: Yes, since the p-value is 0.04 (less than 0.05), we reject the null hypothesis and conclude that there is a seasonal trend in this series. Also, ozone varies a lot by weather season, with the most being in the summer. Even during the day, the most ozone occurs in the daylight hours, due to the nature of the chemical ozone. We would expect to see more ozone during daylight hours, and more ozone during summertime.

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(GaringerOzone.monthly, aes(x = Date, y = `mean(Concentrations_Clean)`)) +
  geom_line() +
  geom_point()+
  labs(x = "Year", y = expression("Concentration in ppm", title="Ozone Concentration overtime"))
```




```
#Ozone_Trend2<- trend::smk.test(GaringerOzone.monthly.ts)
#summary(Ozone_Trend2)
```

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 and the seasonal Mann Kendall test suggest there is a seasonal trend in this dataset. None of the seasons are statistically significant. In terms of the research question, since the p-value is 0.04 (less than 0.05), we reject the null hypothesis and conclude that there is a seasonal trend in this series.

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.noseason <- GaringerOzone.monthly.Decomposed$time.series[,1:3]

#16
no.season.trend <- Kendall::MannKendall(GaringerOzone.monthly.noseason)
summary(no.season.trend)
```

```
## Score =  -3660 , Var(Score) = 5204000
## denominator =  64349.43
## tau = -0.0569, 2-sided pvalue =0.10872
```

Answer: The score of -3660 suggests a overall downward trend. In terms of the research questions, since the p-value is 0.109 we would conclude that the p-value is not statistically signficant, and therefore say that we cannot reject the null hyptothesis that says there has been no change in ozone at these stations in the 2010s.