Assignment 8: Time Series Analysis

Samantha Burch

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., "Salk_A06_GLMs_Week1.Rmd") prior to submission.

The completed exercise is due on Tuesday, March 3 at 1:00 pm.

Set up

1. Set up your session:

- Check your working directory
- Load the tidyverse, lubridate, zoo, and trend packages
- Set your ggplot theme
- 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). Call these GaringerOzone201*, with the star filled in with the appropriate year in each of ten cases.

```
#working directory
getwd()
```

```
## [1] "/Users/samanthaburch/Desktop/Data Analytics/Environmental_Data_Analytics_2020"
```

```
#load
library(tidyverse)
library(lubridate)
library(trend)
library(zoo)
library(ggplot2)

#upload data
EPA2010 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2010_raw.csv")
EPA2011 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2011_raw.csv")
EPA2012 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2012_raw.csv")
EPA2013 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2013_raw.csv")
EPA2014 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2014_raw.csv")
EPA2015 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2015_raw.csv")
EPA2016 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2016_raw.csv")
```

```
EPA2017 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2017_raw.csv")
EPA2018 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2018_raw.csv")
EPA2019 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2019_raw.csv")
```

Wrangle

2. Combine your ten datasets into one dataset called GaringerOzone. Think about whether you should use a join or a row bind.
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 2019-12-13 (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 comine 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.

```
# 2
GaringerOzone <- rbind(EPA2010, EPA2011, EPA2012, EPA2013, EPA2014, EPA2015, EPA2016, EPA2017, EPA2018,

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

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

# 5
GaringerDays <- as.data.frame(seq(as.Date('2010-01-01'), as.Date('2019-12-31'), by = "days"), times = nu

colnames(GaringerDays) <- "Date"

# 6
GaringerOzone <- left_join(GaringerDays, GaringerOzone.Wrangle)
```

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

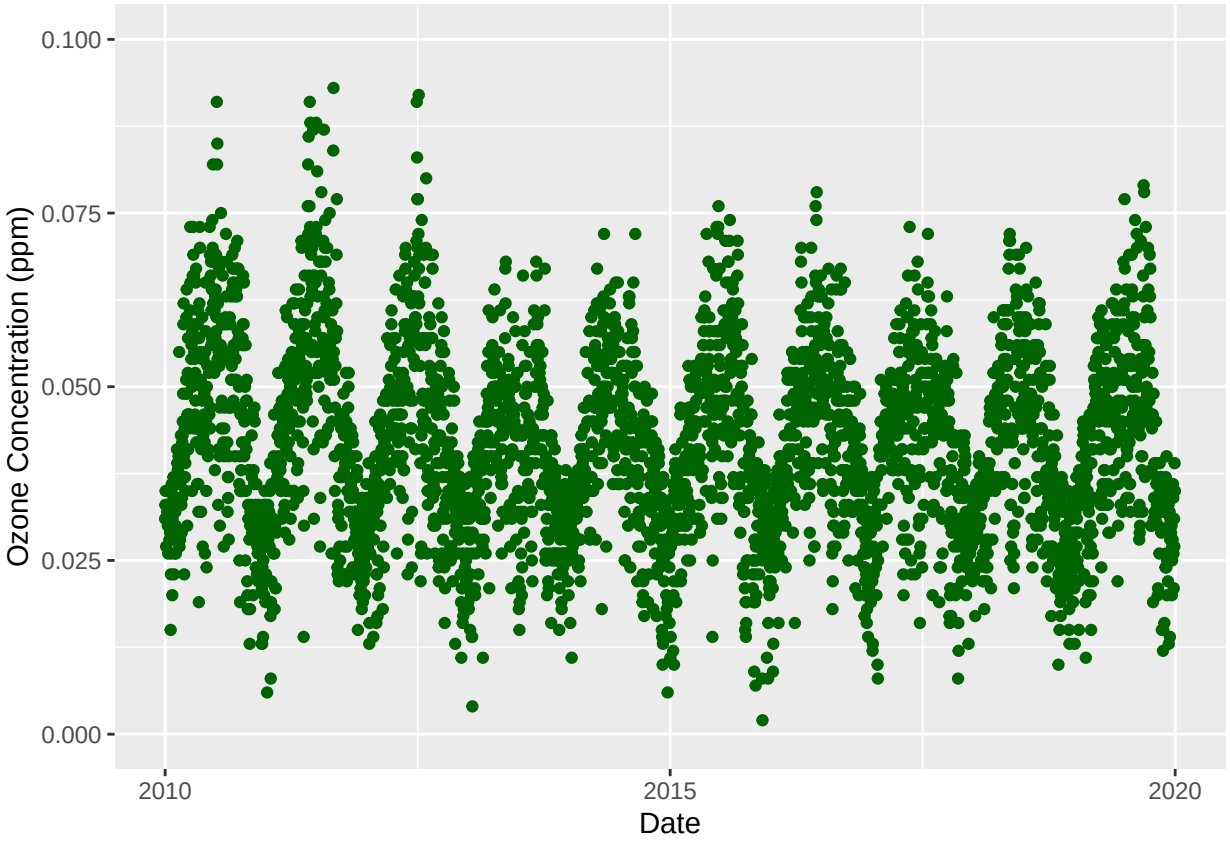
Visualize

7. Create a ggplot depicting ozone concentrations over time. In this case, we will plot actual concentrations in ppm, not AQI values. Format your axes accordingly.

```
GaringerOzone.ggplot <-
ggplot(GaringerOzone, aes(x = Date, y = Daily.Max.8.hour.Ozone.Concentration)) +
  geom_point(color = "dark green") +
  labs(x = "Date", y = "Ozone Concentration (ppm)") +
  ylim(0,.1)

print(GaringerOzone.ggplot)
```

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



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?

   Answer: We didn't use either of these, the 'nearest neighbor approach' nor the 'smooth approach,' because the linear option allows us to better determine the values of the interpolated data on any given date. More specifically, the ozone values vary significantly on a daily basis, and we do not have a large amount of missing data; thus, linear interpolation will work to best fill the few missing data points. For example, spline would not be best, as it would over/under predict lower bounds, and piecewise assumes that any missing data are equal to the measurement made nearest to that date.

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)

10. Generate a time series called GaringerOzone.monthly.ts, with a monthly frequency that specifies the correct start and end dates.

11. Run a time series analysis. In this case the seasonal Mann-Kendall is most appropriate; why is this?

    Answer: This is the most appropriate because it is a nonparametric test that will analyze data for monotonic trends within seasonal data. In this case, the presence of seasonality will imply that our data has varied distributions for different seasons (months). The trend may or may not be linear (upward or downward).

12. To figure out the slope of the trend, run the function `sea.sens.slope` on the time series dataset.
13. Create a plot depicting mean monthly ozone concentrations over time, with both a geom_point and a geom_line layer. No need to add a line for the seasonal Sen's slope; this is difficult to apply to a graph with time as the x axis. Edit your axis labels accordingly.

```
# 8
GaringerOzone$Daily.Max.8.hour.Ozone.Concentration <- na.approx(GaringerOzone$Daily.Max.8.hour.Ozone.Co

# 9
GaringerOzone.monthly <- GaringerOzone %>%
  mutate(Year = year(Date),
         Month = month(Date)) %>%
  group_by(Year, Month) %>%
  summarise(Mean.Monthly = mean(Daily.Max.8.hour.Ozone.Concentration))

GaringerOzone.monthly$Date <- as.Date(paste(GaringerOzone.monthly$Year,
                                            GaringerOzone.monthly$Month,
                                            1, sep="-"),
                                      format = "%Y-%m-%d")

# 10
GaringerOzone.monthly.ts <- ts(GaringerOzone.monthly$Mean.Monthly, frequency = 12,
                          start = c(2010, 01, 01), end = c(2019, 12, 31))

# 11
GaringerOzone.trend <- smk.test(GaringerOzone.monthly.ts)

# 12
sea.sens.slope(GaringerOzone.monthly.ts)
```

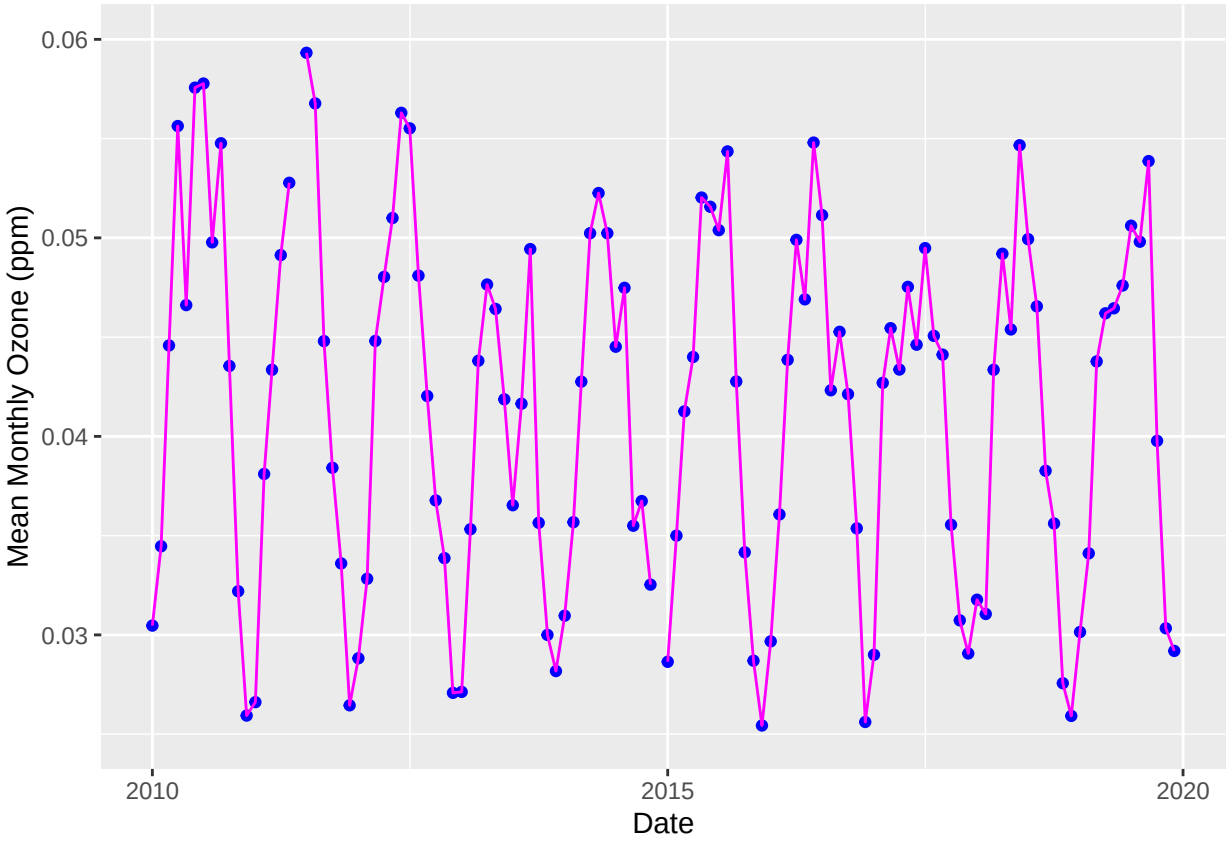
```
## [1] -0.0002044163
```

```
# 13 small negative number - ozone levels are slighlty decreasing over the whole period

GaringerOzone.interpolated <-
ggplot(GaringerOzone.monthly, aes(x = Date, y = Mean.Monthly)) +
  geom_point(color = "blue") +
  geom_line(color = "magenta") +
  labs(x = "Date", y = "Mean Monthly Ozone (ppm)") +
  ylim(0.025, 0.06)

print(GaringerOzone.interpolated)
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



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: We can see that the graph shows that, for Garinger 2010s, there is a negative, signficant monotonic trend for mean monthly ozone concentrations (SMK, z = -1.963, p-value $<$ .05). Over the course of 2010 to 2020, we also note a slight decrease in monthly mean ozone concentrations. The sea sens slope is equal to -0.002; yet, there is not a significant seasonal difference in ozone concentrations across each individual month. Overall, the seasonal component is not significant.