Assignment 7: Time Series Analysis

Devin Domeyer

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
setwd("~/Desktop/Duke/Data Analytics/Environmental_Data_Analytics_2022")
library(tidyverse)
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     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 = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "bottom")
theme_set(mytheme)
```

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
EPA.2010 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2010_raw.csv", stringsAsFactors =
EPA.2011 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2011_raw.csv", stringsAsFactors =
EPA.2012 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2012_raw.csv", stringsAsFactors =
EPA.2013 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2013_raw.csv", stringsAsFactors =
EPA.2014 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2014_raw.csv", stringsAsFactors =
EPA.2015 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2015_raw.csv", stringsAsFactors =
EPA.2016 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2016_raw.csv", stringsAsFactors =
EPA.2017 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2017_raw.csv", stringsAsFactors =
EPA.2018 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2018_raw.csv", stringsAsFactors =
EPA.2019 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2019_raw.csv", stringsAsFactors =

GaringerOzone <- rbind(EPA.2010, EPA.2011, EPA.2012, EPA.2013, EPA.2014,
                       EPA.2015, EPA.2016, EPA.2017, EPA.2018, EPA.2019)
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)
class(GaringerOzone$Date)
```

```
## [1] "Date"
```

```
# 4
GaringerOzone.select <- 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 = "days"))
colnames(Days) <- c("Date")

# 6
GaringerOzone <- left_join(Days, GaringerOzone.select, by = c("Date"))
dim(GaringerOzone)
```

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

Visualize

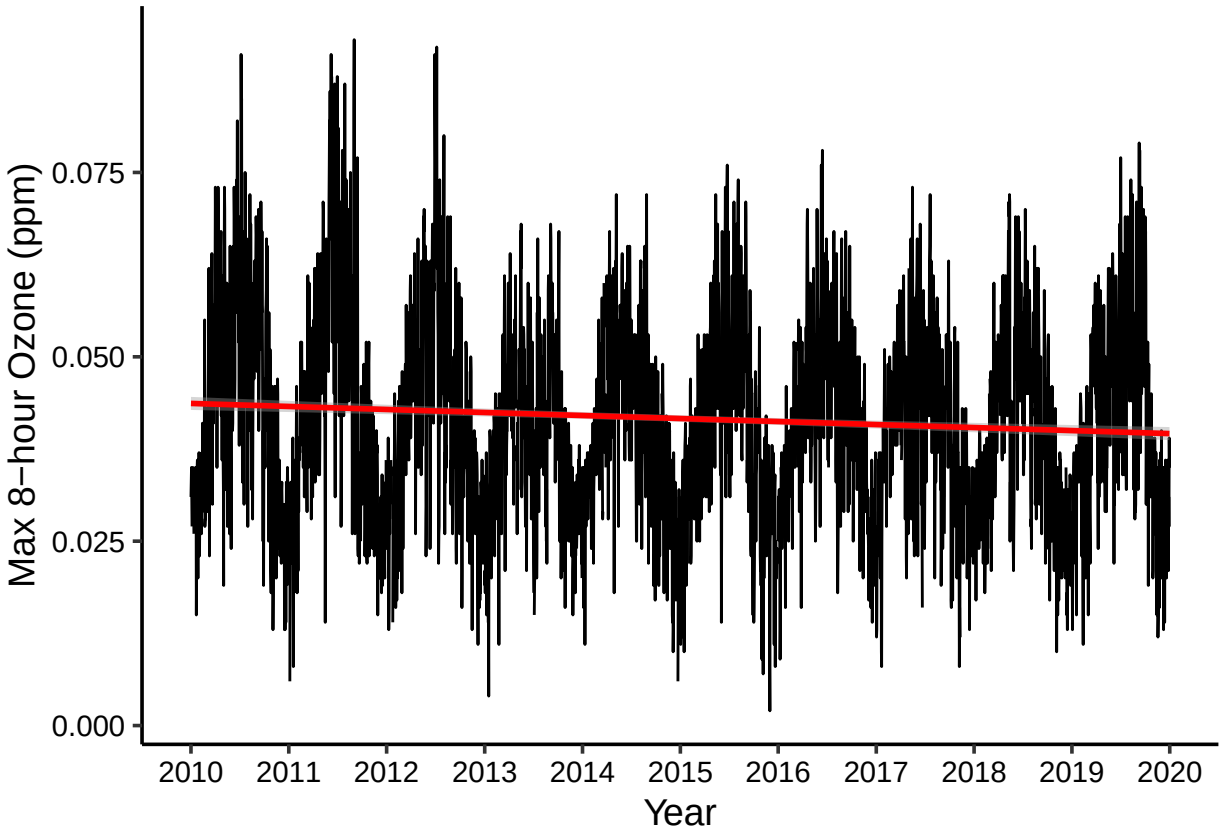
7. Create a line plot depicting ozone concentrations over time. In this case, we will plot actual concentrations in ppm, not AQI values. Format your axes accordingly. Add a smoothed line showing any linear trend of your data. Does your plot suggest a trend in ozone concentration over time?

```
#7

ozoneplot <- ggplot(GaringerOzone, aes(x=Date, y=Daily.Max.8.hour.Ozone.Concentration)) +
  geom_line() +
  geom_smooth(method = lm, color = "red") +
  scale_x_date(date_breaks = "years", date_labels = "%Y") +
  ylab("Max 8-hour Ozone (ppm)") +
  xlab("Year")
print(ozoneplot)
```

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

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



Answer: The plot suggest a very slight downward trend in ozone concentration over time, with definitive seasonal variation.

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

Answer: We didn't use the piecewise interpolation because of the seasonal variability in the data – the nearest neighbor might have a value very different than what the N/A should have represented. We didn't use the spline interpolation because a quadratic function would not fit the data well, as there are multiple vertices observed in the dataset. A linear interpolation is the best fit in this situation.

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.clean %>%
  mutate(month = month(Date), year = year(Date)) %>%
  mutate(month_year = my(paste0(month, "-", year))) %>%
  group_by(month_year) %>%
  summarize(mean.Ozone = mean(Daily.Max.8.hour.Ozone.Concentration))

fd_month <- month(first(GaringerOzone.clean$Date))
fd_year <- year(first(GaringerOzone.clean$Date))

fm_month <- month(first(GaringerOzone.monthly$month_year))
fm_year <- year(first(GaringerOzone.monthly$month_year))
```

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.

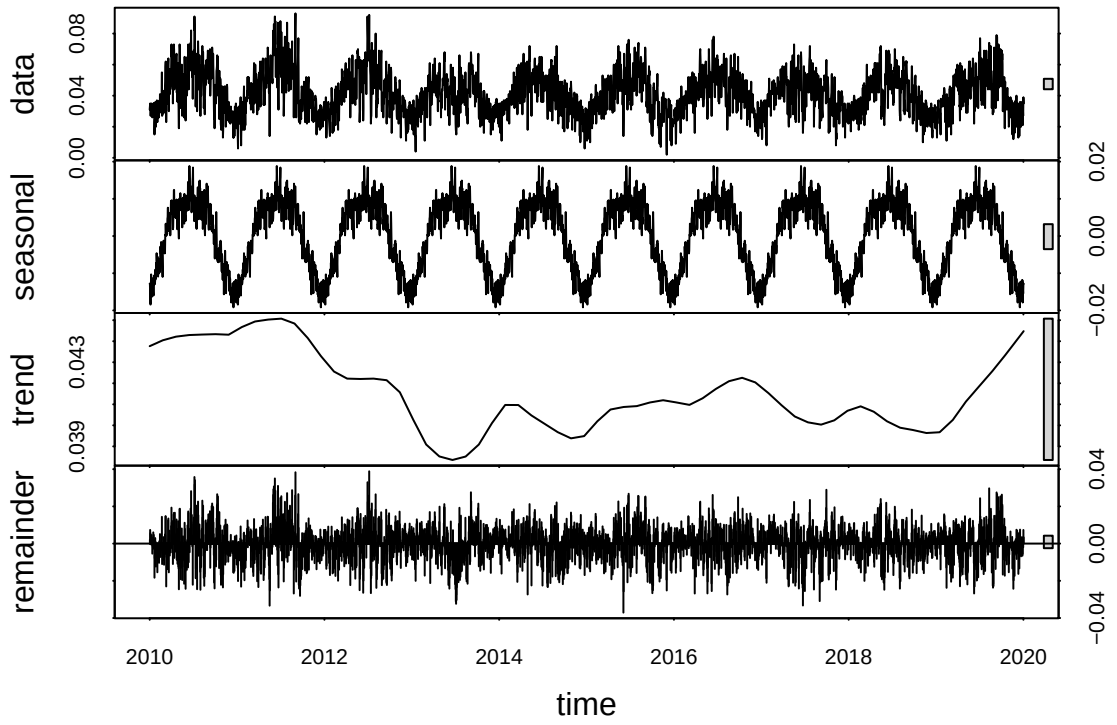
```
#10
GaringerOzone.daily.ts <- ts(GaringerOzone.clean$Daily.Max.8.hour.Ozone.Concentration,
                             start=c(fd_year, fd_month), frequency = 365)

GaringerOzone.monthly.ts <- ts(GaringerOzone.monthly$mean.Ozone, start=c(fm_year, fm_month),
                             frequency = 12)
```

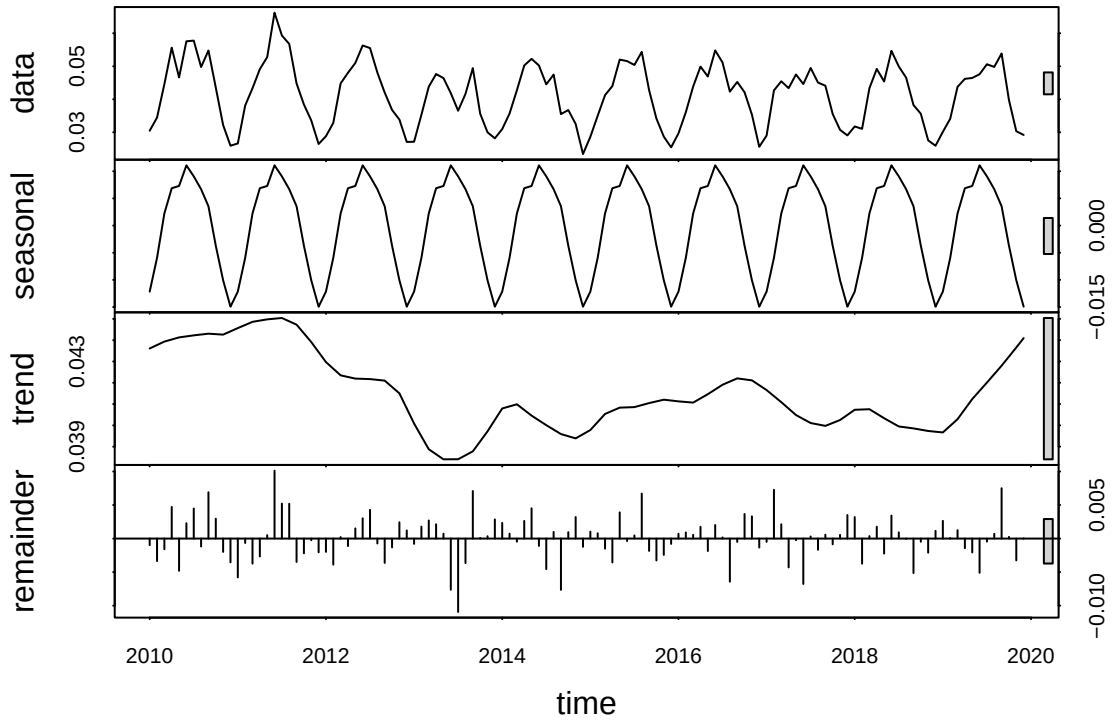
11. Decompose the daily and the monthly time series objects and plot the components using the `plot()` function.

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

plot(Garinger.daily.decomp)
```




```
plot(Garinger.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
Garinger.monthly.trend <- Kendall::SeasonalMannKendall(GaringerOzone.monthly.ts)
summary(Garinger.monthly.trend)
```

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

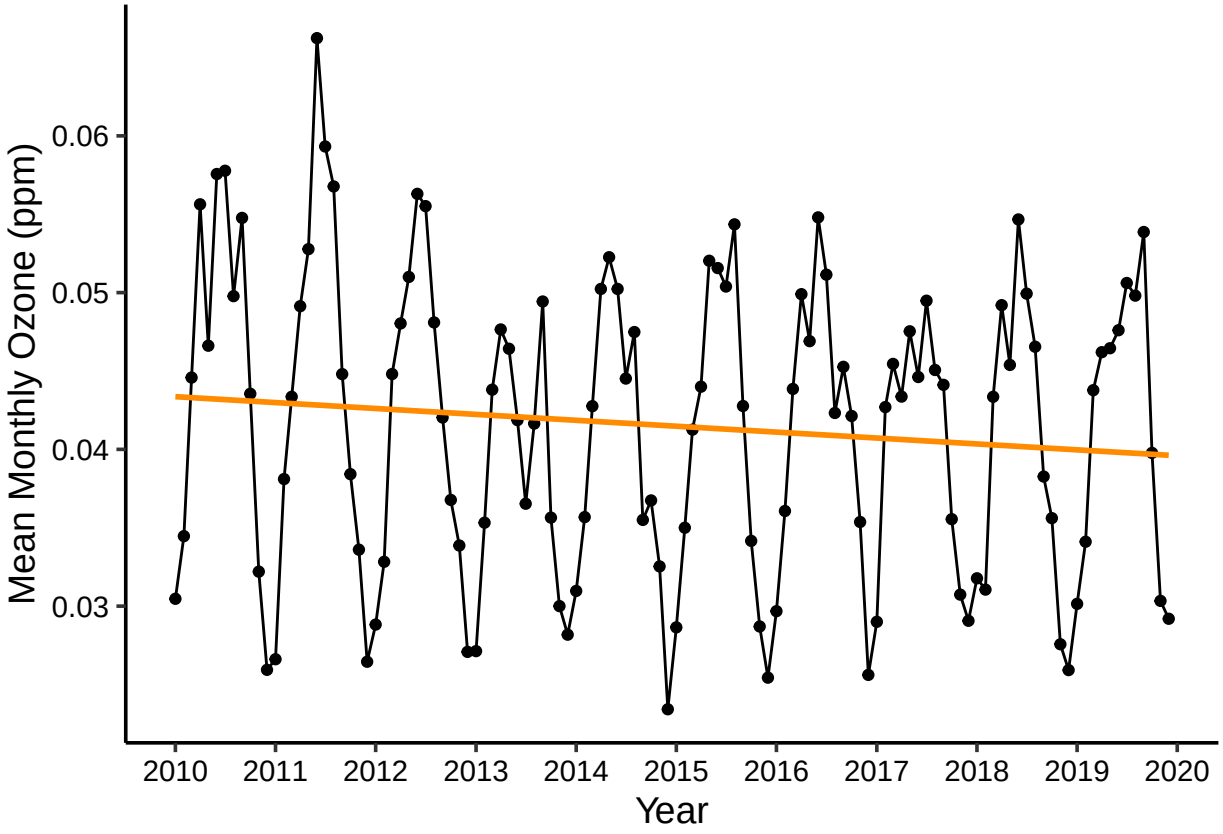
Answer: It is most appropriate because our data is clearly seasonal and non-parametric. The Seasonal Mann-Kendall is the only trend analysis that can handle seasonal data.

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
Garinger.monthly.plot <- ggplot(GaringerOzone.monthly, aes(x=month_year, y=mean.Ozone)) +
  geom_point() +
  geom_line() +
  geom_smooth(method = lm, se = FALSE, color = "darkorange") +
  scale_x_date(date_breaks = "years", date_labels = "%Y") +
  ylab("Mean Monthly Ozone (ppm)") +
  xlab("Year")

print(Garinger.monthly.plot)
```

```
## `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: Ozone concentrations have changed over the 2010s at this station, with a significant inverse relationship between mean monthly ozone and year (S = -77, p = 0.0467).

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
Garinger.Components <- as.data.frame(Garinger.monthly.decomp$time.series[,1:3])
Garinger.non_seasonal <- GaringerOzone.monthly.ts - Garinger.Components$seasonal


#16
Kendall::MannKendall(Garinger.non_seasonal)
```

```
## tau = -0.165, 2-sided pvalue =0.0075402
```

Answer: The result from the non-seasonal trend analysis, once the seasonal variability had been removed from the original Garinger dataset, is far more significant than the seaonsal Mann-Kendall test. The seasonal test has a p-value of 0.047, whereas the non-seasonal test has a p-value of 0.008. This suggests that seasonal variation influences the data toward a more stationary trend, but that there is still a significant downward sloping linear trend to the data.