Assignment 8: Time Series Analysis

Lauren Ng

Fall 2024

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on generalized linear models.

Directions

1. Rename this file `<FirstLast>_A08_TimeSeries.Rmd` (replacing `<FirstLast>` with your first and last name).
2. Change "Student Name" on line 3 (above) with your name.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.

Set up

1. Set up your session:

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

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(lubridate)
library(zoo)
```

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

```
library(trend)

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

```
#1
Garinger.2010 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2010_raw.csv",
                          stringsAsFactors = TRUE)
Garinger.2011 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2011_raw.csv",
                          stringsAsFactors = TRUE)
Garinger.2012 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2012_raw.csv",
                          stringsAsFactors = TRUE)
Garinger.2013 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2013_raw.csv",
                          stringsAsFactors = TRUE)
Garinger.2014 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2014_raw.csv",
                          stringsAsFactors = TRUE)
Garinger.2015 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2015_raw.csv",
                          stringsAsFactors = TRUE)
Garinger.2016 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2016_raw.csv",
                          stringsAsFactors = TRUE)
Garinger.2017 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2017_raw.csv",
                          stringsAsFactors = TRUE)
Garinger.2018 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2018_raw.csv",
                          stringsAsFactors = TRUE)
Garinger.2019 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2019_raw.csv",
                          stringsAsFactors = TRUE)

GaringerOzone <- rbind(Garinger.2010,Garinger.2011,Garinger.2012,Garinger.2013,
                       Garinger.2014,Garinger.2015,Garinger.2016,Garinger.2017,
                       Garinger.2018,Garinger.2019)
```

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

# 4

GaringerOzone_Wrangled <- 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_Wrangled, by = "Date")
```

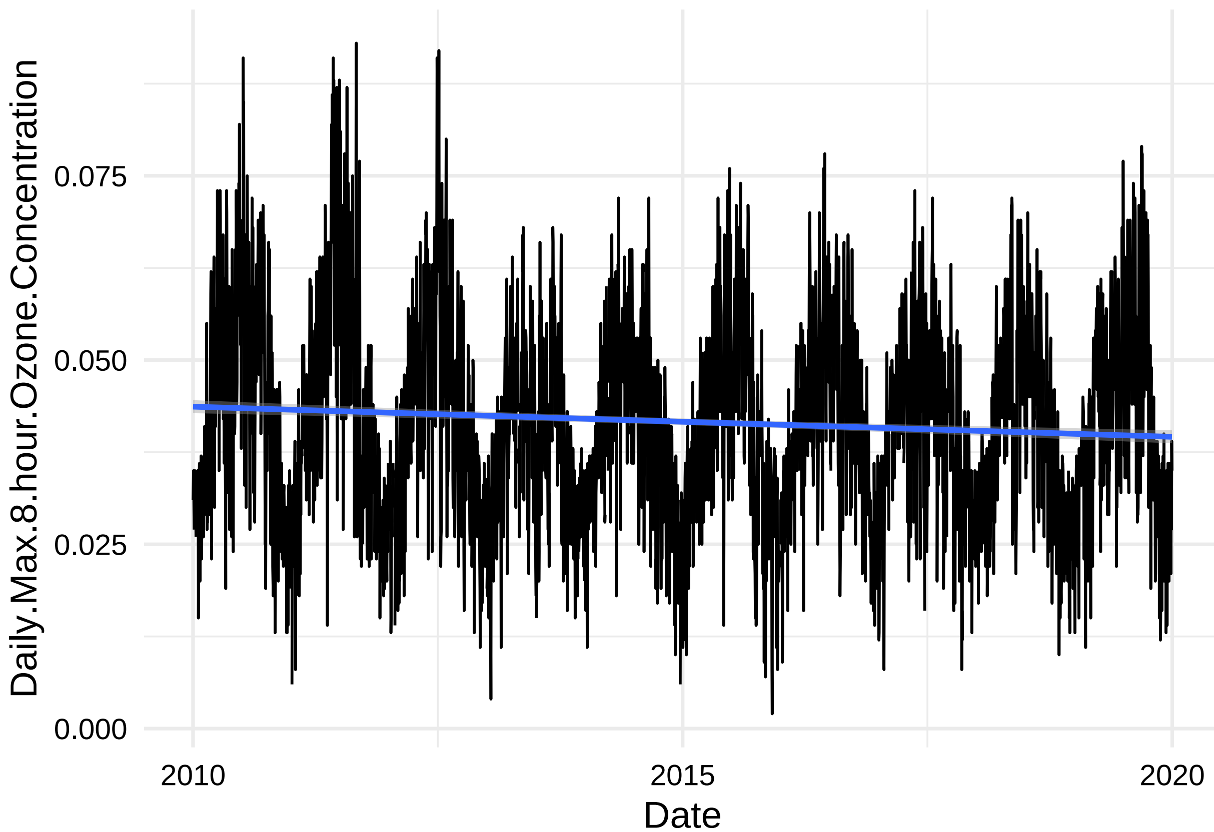
Visualize

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
GaringerOzonePlot <- GaringerOzone %>%
  ggplot(aes(x=Date, y=Daily.Max.8.hour.Ozone.Concentration))+
  geom_line()+
  geom_smooth(method = "lm")

GaringerOzonePlot
```

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

```
## Warning: Removed 63 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



Answer: The plot shows a weak negative correlation but it is too early to draw any conclusions.

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

Answer: We didn't use piecewise constant because that would make the missing data equal to its nearest neighbor. And since we saw in the plot that there was a trend in the ozone concentrations over time, linear interpolation is a better choice to create smoother looking data. The spline interpolation would use quadratic function rather than a straight line, and we see that our data is more linear, so that would not be a good fit.

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)

```r
#9
GaringerOzone.monthly <- Garinger.Clean %>%
  mutate(Year = as.numeric(format(Date, "%Y")),
    Month = as.numeric(format(Date, "%m"))) %>%
  group_by(Year, Month) %>%
  summarize(Mean_Ozone = mean(Daily.Max.8.hour.Ozone.Concentration.clean, na.rm = TRUE))
```

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

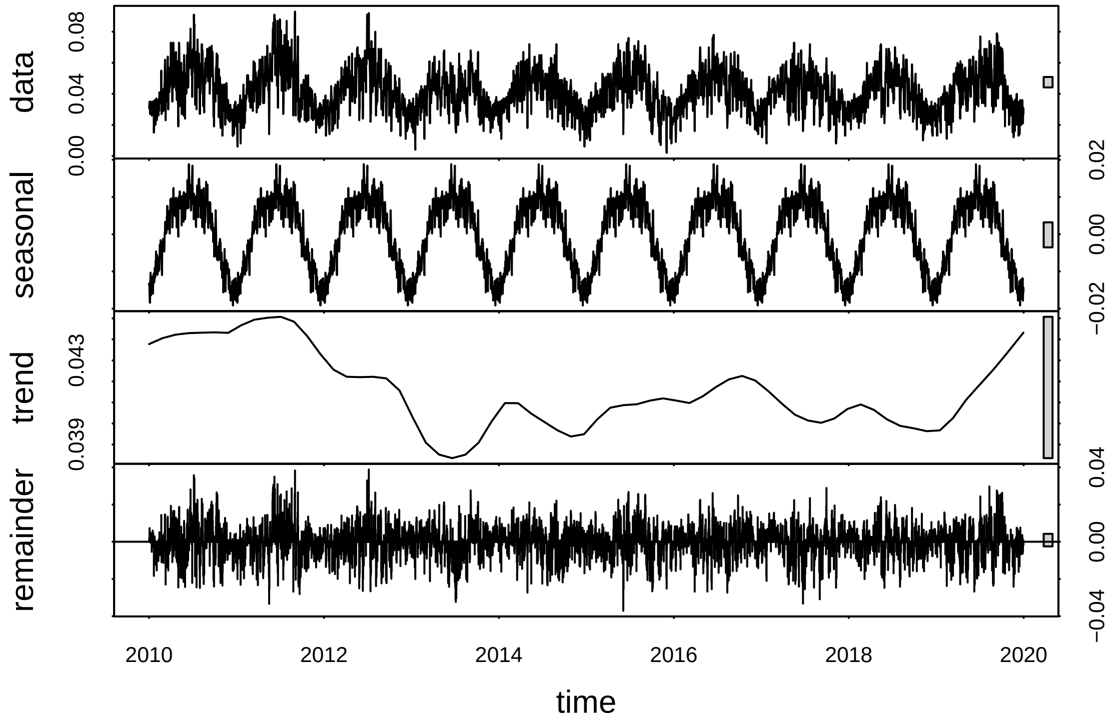
```r
GaringerOzone.monthly <- GaringerOzone.monthly %>%
  mutate(Date = as.Date(paste(Year, Month, "01", sep = "-")))
```

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.

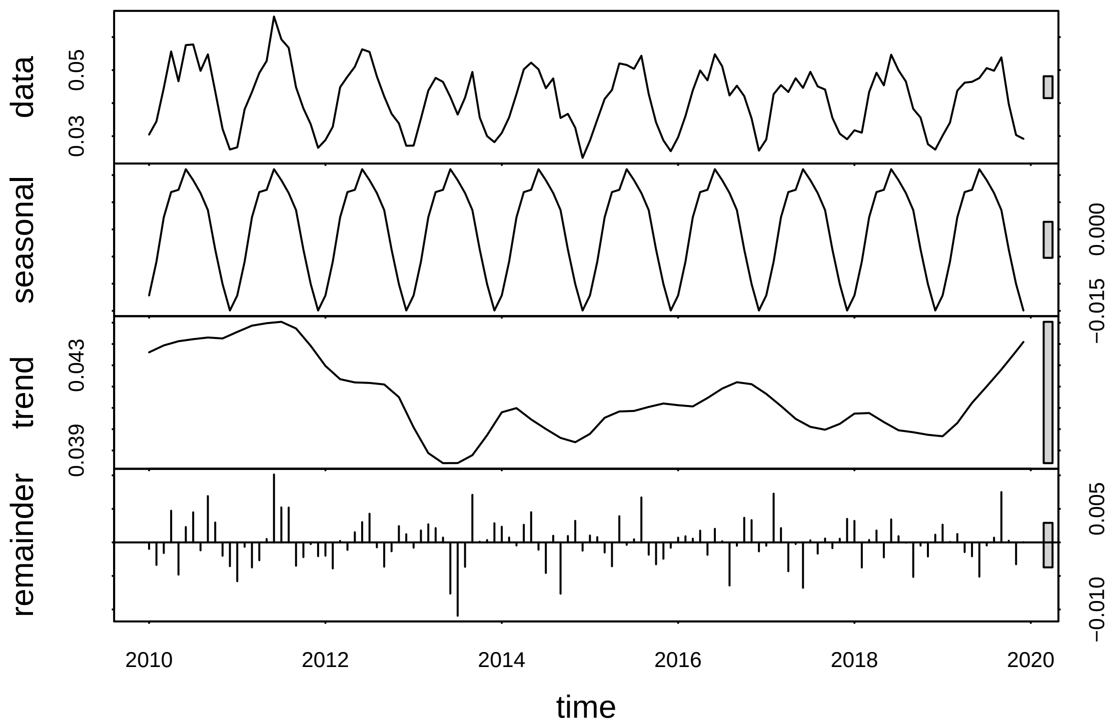
```r
#10
GaringerOzone.daily.ts <- ts(
  Garinger.Clean$Daily.Max.8.hour.Ozone.Concentration.clean,
  start = c(2010, 1), end = c(2019, 365),frequency = 365)

GaringerOzone.monthly.ts <- ts(
  GaringerOzone.monthly$Mean_Ozone,
  start = c(2010, 1), end = c(2019, 12),frequency = 12)
```

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

```r
#11
Garinger_daily__decomp <- stl(GaringerOzone.daily.ts,s.window = "periodic")
plot(Garinger_daily__decomp)
```




```
Garinger_monthly_decomp <- stl(GaringerOzone.monthly.ts,s.window = "periodic")
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.trend1 <- Kendall::SeasonalMannKendall(GaringerOzone.monthly.ts)

summary(Garinger.trend1)
```

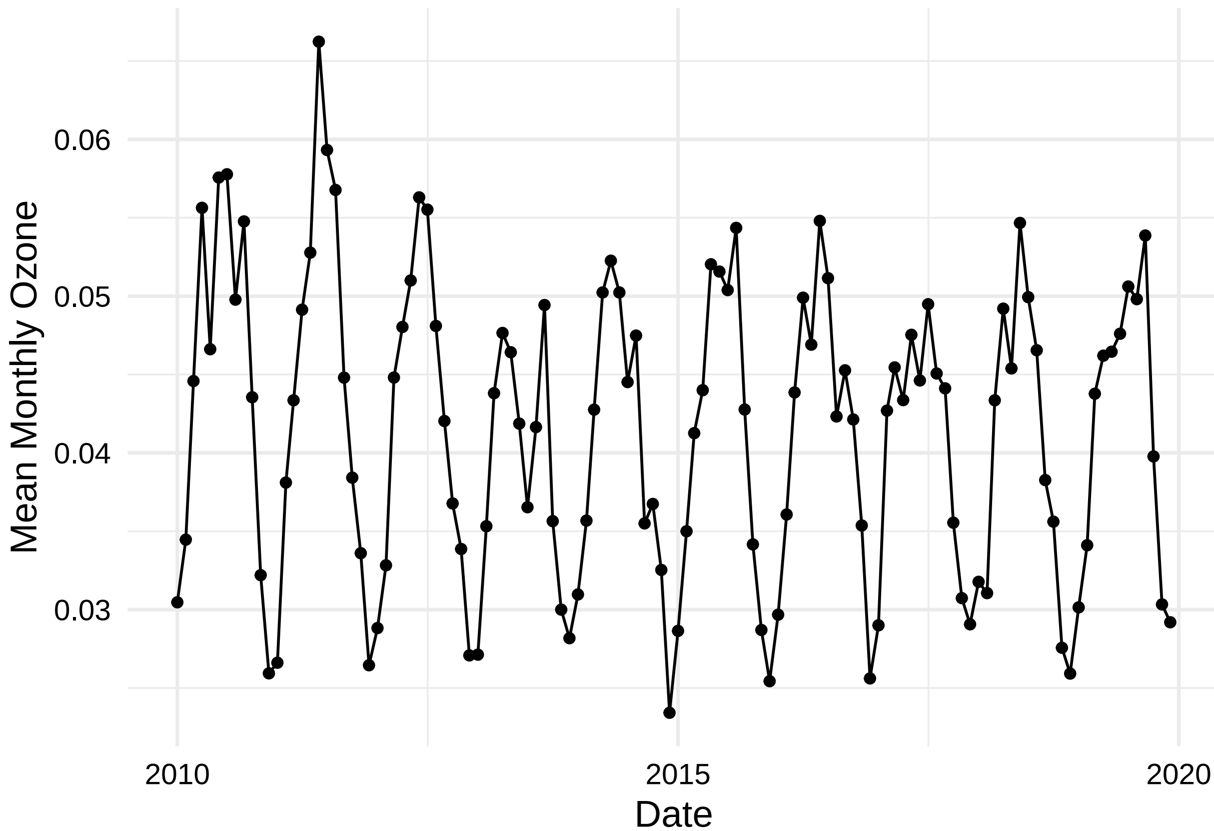
```
## Score =  -77 , Var(Score) = 1499
## denominator =  539.4972
## tau = -0.143, 2-sided pvalue =0.046724
```

Answer: Seasonal Mann-Kendall is most appropriate, because as we can see from the decomposed plot, there is seasonality to the data. We might want to see how ozone has changed over time while incorporating the seasonal component.

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
MeanOzonePlot <- GaringerOzone.monthly %>%
  ggplot(aes(x=Date, y=Mean_Ozone))+
  geom_point()+
  geom_line()+
```

```
  ylab("Mean Monthly Ozone")

MeanOzonePlot
```



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 wanted to know if ozone concentrations varied year over year. The results of this test show a negative trend with a score of -88, indicating that there is a downward trend in ozone concentration over the year. There is high variance which means ozone concentration fluctuates. Teh tau of -0.163 shows a weak negative trend over time. The p value is less than 0.05, meaning the results are statistically significant. This means that there is a temporal trend of ozone concenration, and that ozone concentration is affected by seasonal factors.

    (Score = -88 , Var(Score) = 1498 denominator = 538.9944 tau = -0.163, 2-sided pvalue =0.022986)

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.

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

GaringerOzone.monthly.adjusted <- GaringerOzone.monthly.ts - monthly_decomposed$time.series[,1]

#16

Garinger.trend2 <- Kendall::MannKendall(GaringerOzone.monthly.adjusted)

summary(Garinger.trend2)
```

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

> Answer: These results show an even greater negative correlation and with a much smaller p-value, greater variance, and a slightly more negative tau showing negative trend over time after the seasonal component was removed. This means that although seasonality is one influence, it is not the greatest influence on ozone concentration, and there are likely other factors which we did not anaylze.