Assignment 8: Time Series Analysis

Lucy Wang

Fall 2023

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

```
# Check working directroy
library(here)
```

```
## here() starts at /Users/lucywang/Documents/EDE_Fall2023
```

```
here()
```

```
## [1] "/Users/lucywang/Documents/EDE_Fall2023"
```

```
# load packages
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## 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 errors
```

```
library(lubridate)
#install.packages("trend")
library(trend)
#install.packages("zoo")
library(zoo)
```

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

```
#install.packages("Kendall")
library(Kendall)
#install.packages("tseries")
library(tseries)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
# set my theme
mytheme <- theme_classic(base_size = 11) +
  theme(
    axis.text = element_text(color = "black"),
    plot.title = element_text(hjust = 0.5),
    legend.position = "top",
    legend.title = element_blank()
  )
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
file_list <-
  list.files(
    '~/Documents/EDE_Fall2023/Data/Raw/Ozone_TimeSeries',
    pattern = "*.csv",
    recursive = TRUE,
    full.names = TRUE
  )
```

```
csv_reader <- function(i) {
  file_i <- read.csv(file_list[i], stringsAsFactors = TRUE)
}

GaringerOzone <- data.frame()

for (i in 1:length(file_list)) {
  GaringerOzone <- rbind(GaringerOzone, csv_reader(i))
}
```

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

# 5
Days <- as.data.frame(seq(
  from = as.Date("2010-01-01"),
  to = as.Date("2019-12-31"),
  by = "day"
))
names(Days)[1] <- 'Date'

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

```
## Joining with ‘by = join_by(Date)‘
```

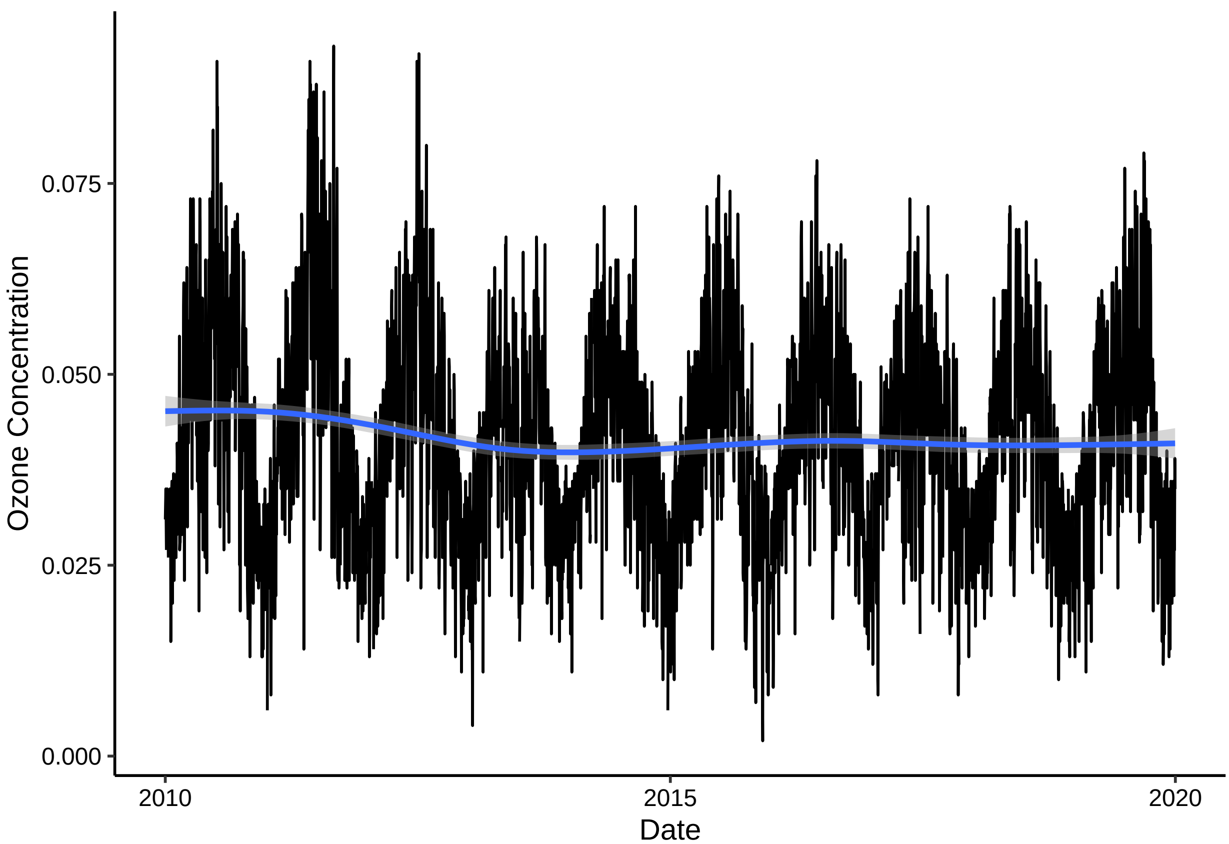
Visualize

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?

```r
#7
ggplot(GaringerOzone,
       aes(x = Date, y = Daily.Max.8.hour.Ozone.Concentration)) +
  geom_line() +
  geom_smooth() +
  ylab("Ozone Concentration")
```

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

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



Answer: The plot suggests a seasonal fluctuation of ozone concentrations and a overall stable variation overtime.

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?

```r
#8
#  Use a linear interpolation to fill NAs
GaringerOzone <- GaringerOzone %>%
  mutate(
    Daily.Max.8.hour.Ozone.Concentration =
      zoo::na.approx(Daily.Max.8.hour.Ozone.Concentration)
  )
```

> Answer: Piecewise fills data with the neariest data, and spline uses a quatratic function. In this case, we want to know the overall trend which should be presented as a linear relationship, so the data filled should follow the continuous 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)

```r
#9
GaringerOzone.monthly <- GaringerOzone %>%
  mutate(Year = year(Date)) %>%
  mutate(Month = month(Date)) %>%
  group_by(Year, Month) %>%
  summarize(Daily.Max.8.hour.Ozone.Concentration =
              mean(Daily.Max.8.hour.Ozone.Concentration)) %>%
  mutate(Date = my(paste0(Month, "-", Year)))
```

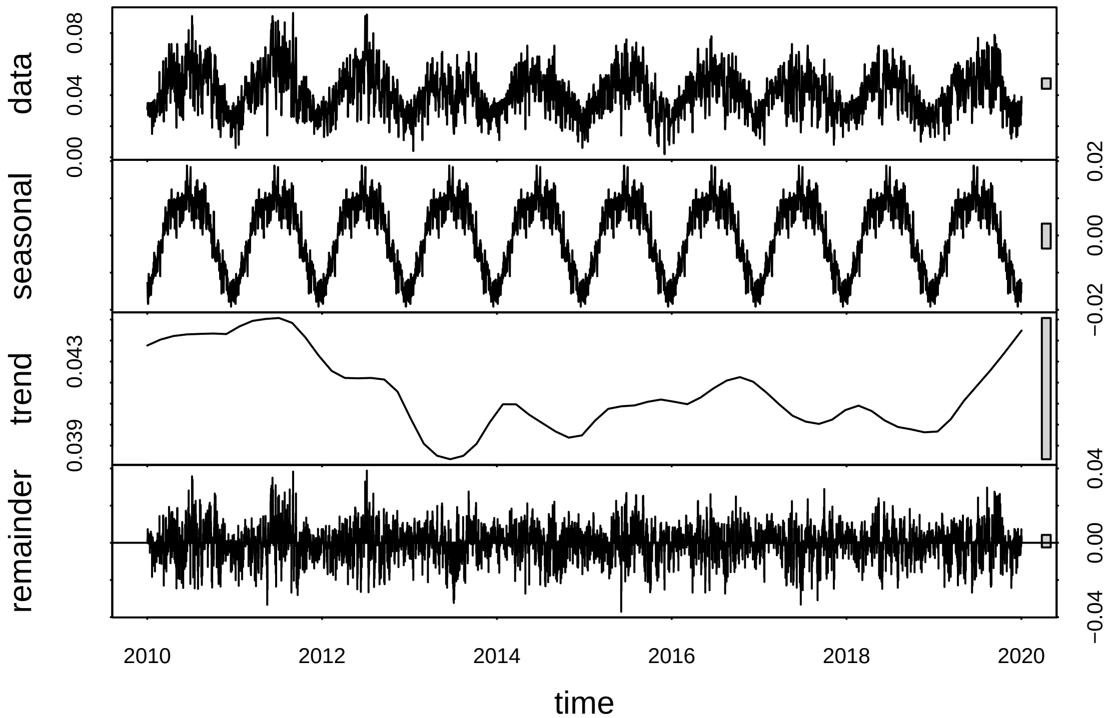
```
## `summarise()` has grouped output by 'Year'. You can override using the
## `.groups` argument.
```

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.

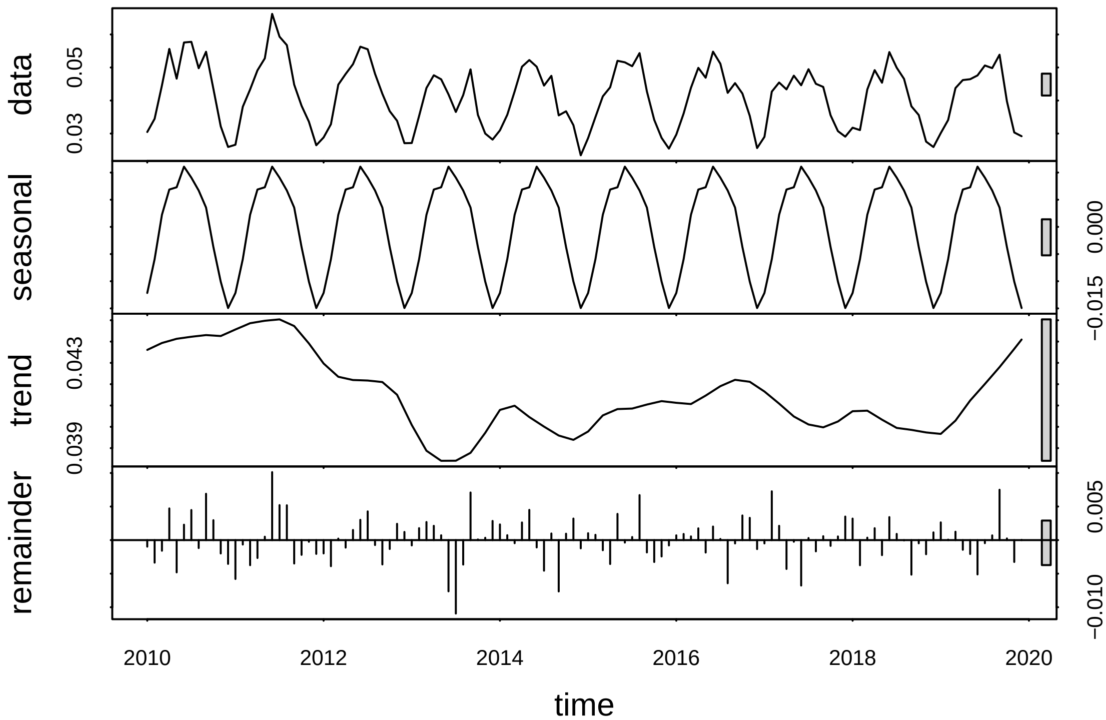
```r
#10
# define start month and year
f_month = month(first(GaringerOzone$Date))
f_year = year(first(GaringerOzone$Date))

# Generate ts based on daily observations
GaringerOzone.daily.ts <-
  ts(
    GaringerOzone$Daily.Max.8.hour.Ozone.Concentration,
    start = c(f_year, f_month),
    frequency = 365
  )

# Generate ts based on monthly observations
GaringerOzone.monthly.ts <- ts(
  GaringerOzone.monthly$Daily.Max.8.hour.Ozone.Concentration,
  start = c(f_year, f_month),
  frequency = 12
)
```

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

```
#11
# Decompose the daily time series
GaringerOzone.daily.decomposed <-
  stl(GaringerOzone.daily.ts, s.window = 'periodic')
plot(GaringerOzone.daily.decomposed)
```




```
# Decompose the monthly time series
GaringerOzone.monthly.decomposed <-
  stl(GaringerOzone.monthly.ts, s.window = 'periodic')
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?

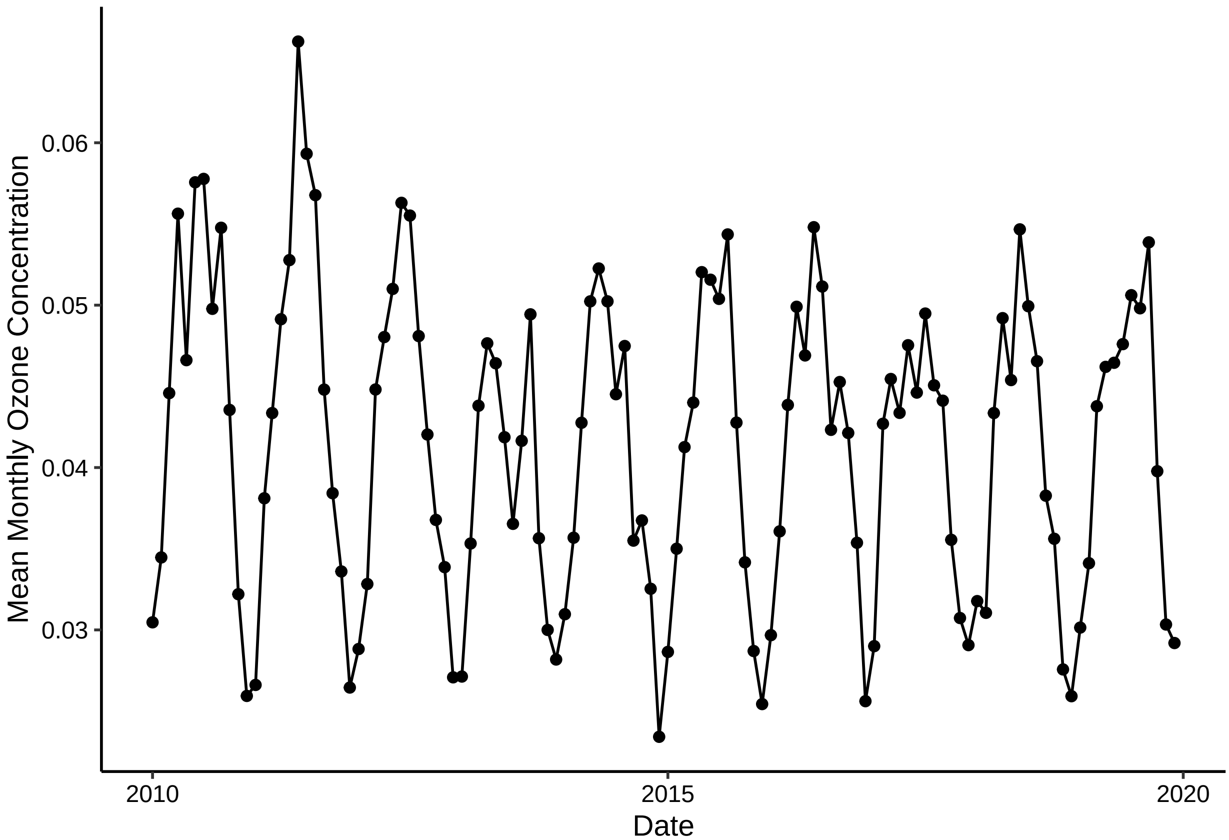
```
#12
GaringerOzone.monthly.trend <-
  Kendall::SeasonalMannKendall(GaringerOzone.monthly.ts)
summary(GaringerOzone.monthly.trend)
```

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

Answer: Because the seasonal Man-Kendall analysis is appropriate for data that has clear seasonal patterns.

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 = Daily.Max.8.hour.Ozone.Concentration)) +
  geom_point() +
  geom_line() +
  ylab('Mean Monthly Ozone Concentration')
```



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: According to the seasonal Mann-Kendall test, there is a statistically significant monotonic trend in Ozone concentrations over the 2010s at this station (p-value = 0.0467 < 0.05). From the graphs, we can see clear seasonal patterns of Ozone concentration.

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
# Extract components
GaringerOzone.monthly_Components <-
  as.data.frame(GaringerOzone.monthly.decomposed$time.series[, 1:3])

# Subtract the seasonal component
GaringerOzone.monthly.NonSeasonal.ts <-
  GaringerOzone.monthly.ts -
  GaringerOzone.monthly_Components$seasonal

#16
GaringerOzone.monthly.NonSeasonalTrend <-
```

```
  Kendall::MannKendall(GaringerOzone.monthly.NonSeasonal.ts)
summary(GaringerOzone.monthly.NonSeasonalTrend)
```

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

> Answer: The Mann Kendall test on the non-seasonal Ozone monthly series suggests a more statistically significant trend than the Seasonal Mann Kendall on the complete series (p-value for the non-seasonal Ozone monthly series is 0.0075, lower than 0.0467).