Assignment 8: Time Series Analysis

Ye Khaung Oo

Spring 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

```
#1
getwd()
```

```
## [1] "/Users/yekhaungoo/Library/CloudStorage/OneDrive-Personal/EDA Class"
```

```
library(tidyverse);library(lubridate);library(zoo);
library(trend);library(ggthemes);library(ggplot2);
library(dplyr); library(Kendall)

yko_theme <- theme_clean() +
  theme(
    axis.title = element_text(size = 12),
    axis.text = element_text(size = 10),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()
  )
```

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

data_folder <- "./Data/Raw/Ozone_TimeSeries"
Ozone_files <- list.files(data_folder, full.names = TRUE)
Ozone_df <- lapply(Ozone_files, read.csv)
GaringerOzone <- do.call(rbind, Ozone_df)
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$Date <- as.Date(GaringerOzone$Date, format = "%m/%d/%Y")

# 4
GaringerOzone <- 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, by = "Date")
```

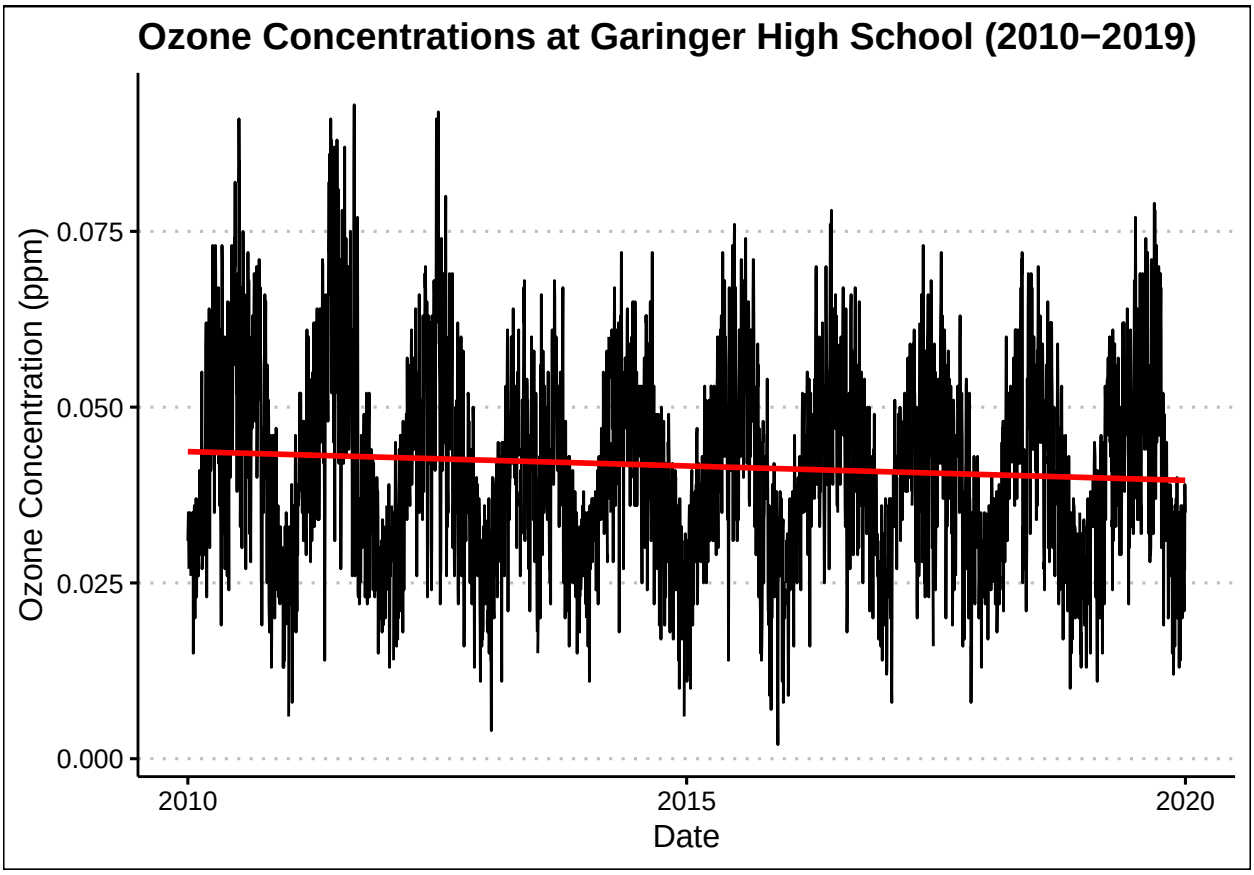
Visualize

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(method = "lm", se = FALSE, color = "red") +
  labs(x = "Date", y = "Ozone Concentration (ppm)",
       title = "Ozone Concentrations at Garinger High School (2010-2019)") +
  yko_theme
```

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

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



Answer: The plot shows a negative trend in Ozone concentration over the years.

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?

```
knitr::opts_chunk$set(tidy.opts=list(width.cutoff=60), tidy=TRUE)

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

Answer: Perhaps to avoid the 'overshoot' problem of piecewise constant or spline interpolation. There is no derivative involved in the ozone concentration so linear interpolation might be most appropriate.

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)

```
# knitr::opts_chunk$set(tidy.opts=list(width.cutoff=60),
# tidy=TRUE)

# 9
GaringerOzone <- GaringerOzone %>%
    mutate(Year = year(Date), Month = month(Date))

GaringerOzone.monthly <- GaringerOzone %>%
    group_by(Year, Month) %>%
    summarise(Mean.Ozone.Concentration = mean(Daily.Max.8.hour.Ozone.Concentration,
        na.rm = TRUE))
```

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

```
GaringerOzone.monthly$Date <- as.Date(paste(GaringerOzone.monthly$Year,
    GaringerOzone.monthly$Month, "01", sep = "-"))

# the following code is to answer Question 14
min(GaringerOzone$Daily.Max.8.hour.Ozone.Concentration)
```

```
## [1] 0.002
```

```
max(GaringerOzone$Daily.Max.8.hour.Ozone.Concentration)
```

```
## [1] 0.093
```

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.

```
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = FALSE)

# 10 First Time Series - Daily Ozone
min(GaringerOzone$Date)
```

```
## [1] "2010-01-01"
```

```
max(GaringerOzone$Date)
```

```
## [1] "2019-12-31"
```

```
Ozone.daily.ts <- ts(GaringerOzone$Daily.Max.8.hour.Ozone.Concentration,
    start = c(year(min(GaringerOzone$Date)), month(min(GaringerOzone$Date))),
    end = c(year(max(GaringerOzone$Date)), month(max(GaringerOzone$Date))),
    frequency = 365)

# Second Time Series - Monthly Ozone
min(GaringerOzone.monthly$Date)
```

```
## [1] "2010-01-01"
```

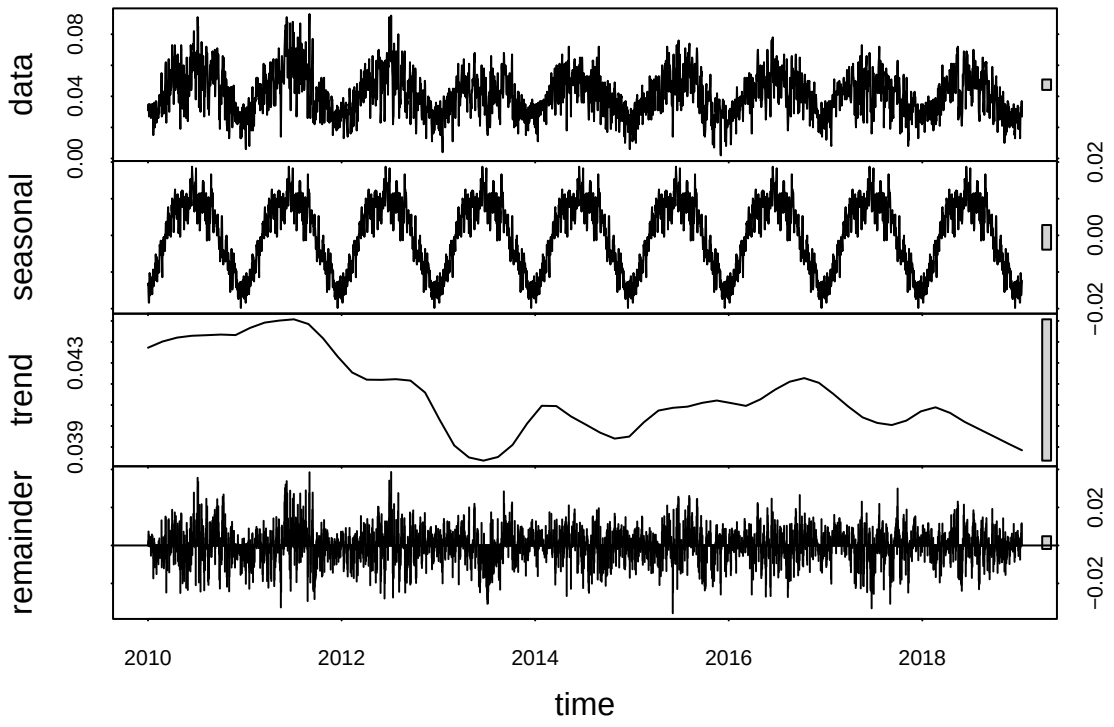
```
max(GaringerOzone.monthly$Date)
```

```
## [1] "2019-12-01"
```

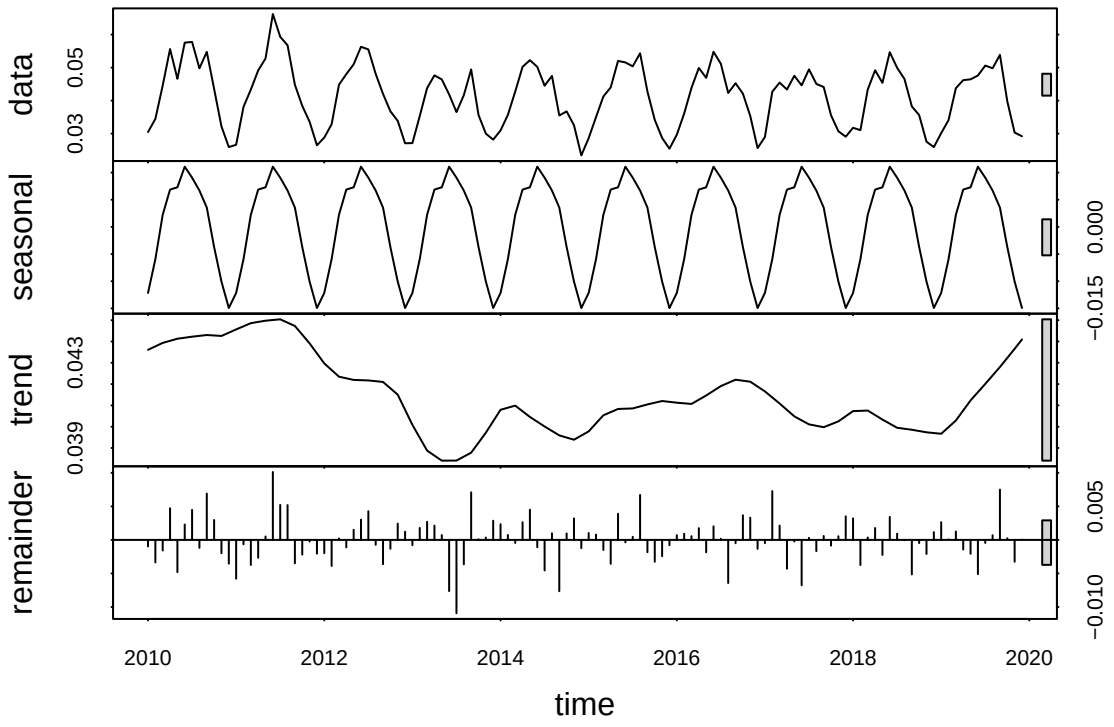
```
Ozone.monthly.ts <- ts(GaringerOzone.monthly$Mean.Ozone.Concentration,
    start = c(year(min(GaringerOzone.monthly$Date)), month(min(GaringerOzone.monthly$Date))),
    end = c(year(max(GaringerOzone.monthly$Date)), month(max(GaringerOzone.monthly$Date))),
    frequency = 12)
```

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

```
#11
Ozone.daily.decomposed <- stl(Ozone.daily.ts, s.window = "periodic")
plot(Ozone.daily.decomposed)
```




```
Ozone.monthly.decomposed <- stl(Ozone.monthly.ts, s.window = "periodic")
plot(Ozone.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?

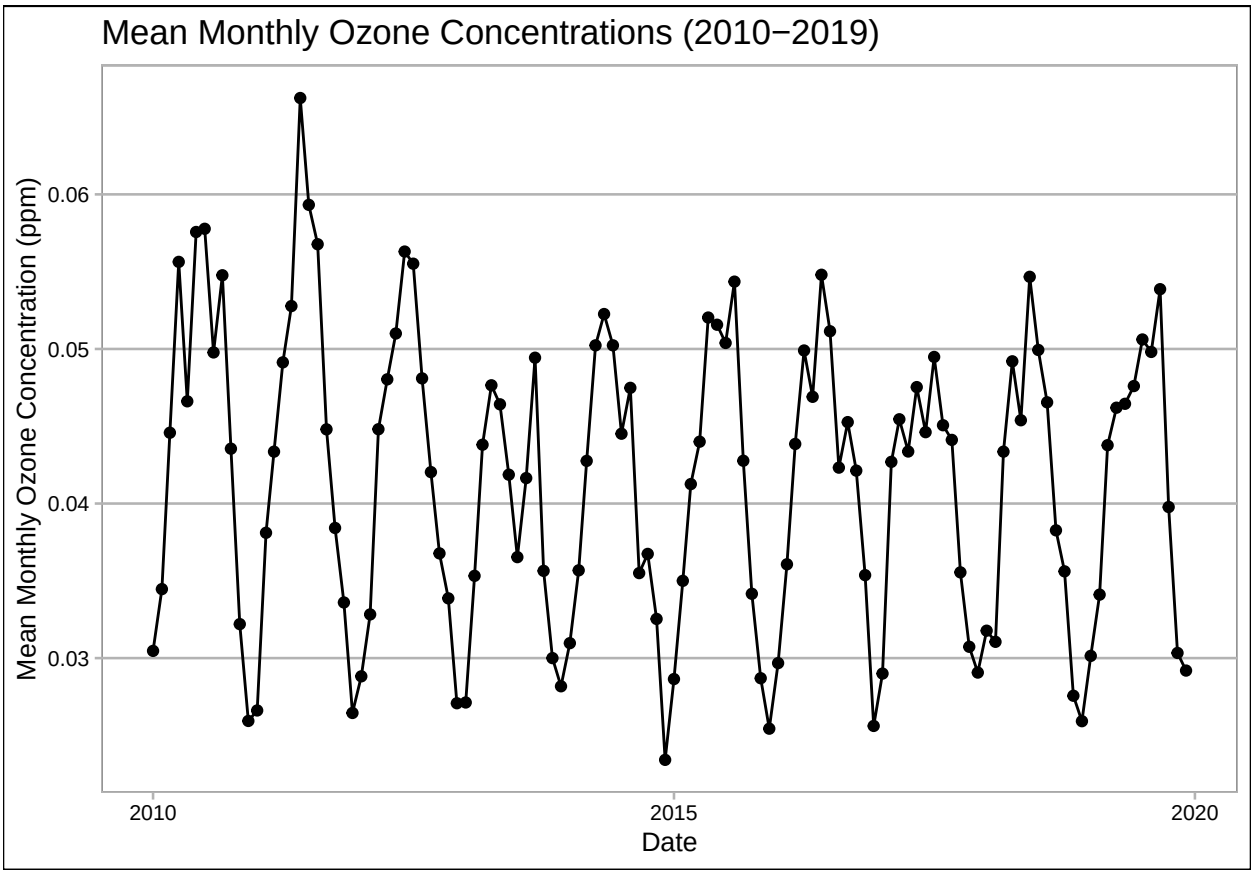
```
#12
Ozone.SMKtest <- smk.test(Ozone.monthly.ts)
Ozone.SMKtest
```

```
##
##  Seasonal Mann-Kendall trend test (Hirsch-Slack test)
##
## data:  Ozone.monthly.ts
## z = -1.963, p-value = 0.04965
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##    S varS
##  -77 1499
```

Answer: The Ozone concentration seems to vary seasonally. Therefore, Seasonal Mann-Kendall test is appropriate since the test takes seasonal patterns into account.

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.

```r
# 13
ggplot(GaringerOzone.monthly, aes(x = Date, y = Mean.Ozone.Concentration)) +
  geom_point() +
  geom_line() +
  labs(x = "Date", y = "Mean Monthly Ozone Concentration (ppm)",
       title = "Mean Monthly Ozone Concentrations (2010-2019)") +
  theme_calc()
```



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 at this station have changed over the 2010s. The negative Z-score indicates a decreasing trend, and the P-value is statistically significant. The Ozone concentration (ppm) ranges from 0.002 to 0.093 during the 2010s, with seasonal variations.

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
Ozone_NonSeasonal_df <- as.data.frame(Ozone.monthly.decomposed$time.series[,
    1:3])
Ozone_NonSeasonal_ts <- Ozone.monthly.ts - Ozone_NonSeasonal_df$seasonal


# 16
Ozone_NonSeasonal_Test <- mk.test(Ozone_NonSeasonal_ts)
Ozone_NonSeasonal_Test
```

```
##
##  Mann-Kendall trend test
##
## data:  Ozone_NonSeasonal_ts
## z = -2.672, n = 120, p-value = 0.00754
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S          varS           tau
## -1.179000e+03  1.943657e+05 -1.651376e-01
```

Answer: Mann-Kendall trend test shows statistical significance and a negative Z-score which indicates a negative trend in Ozone concentration. This result is consistent with the seasonal SMK test done in Q12.