Assignment 8: Time Series Analysis

Jiahuan Li

Spring 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

```
#1
library(here)
library(tidyverse)
library(lubridate)
library(zoo)
library(trend)
library(ggplot2)
library(dplyr)
library(Kendall)

mytheme <- theme_classic(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.

```
#2
files <- list.files(here("Data","Raw","Ozone_TimeSeries"), pattern = "*.csv", full.names = TRUE)
df_list <- lapply(files, read.csv)
GaringerOzone <- bind_rows(df_list)
```

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 <- select(GaringerOzone, 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)
```

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

```
dim(GaringerOzone)
```

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

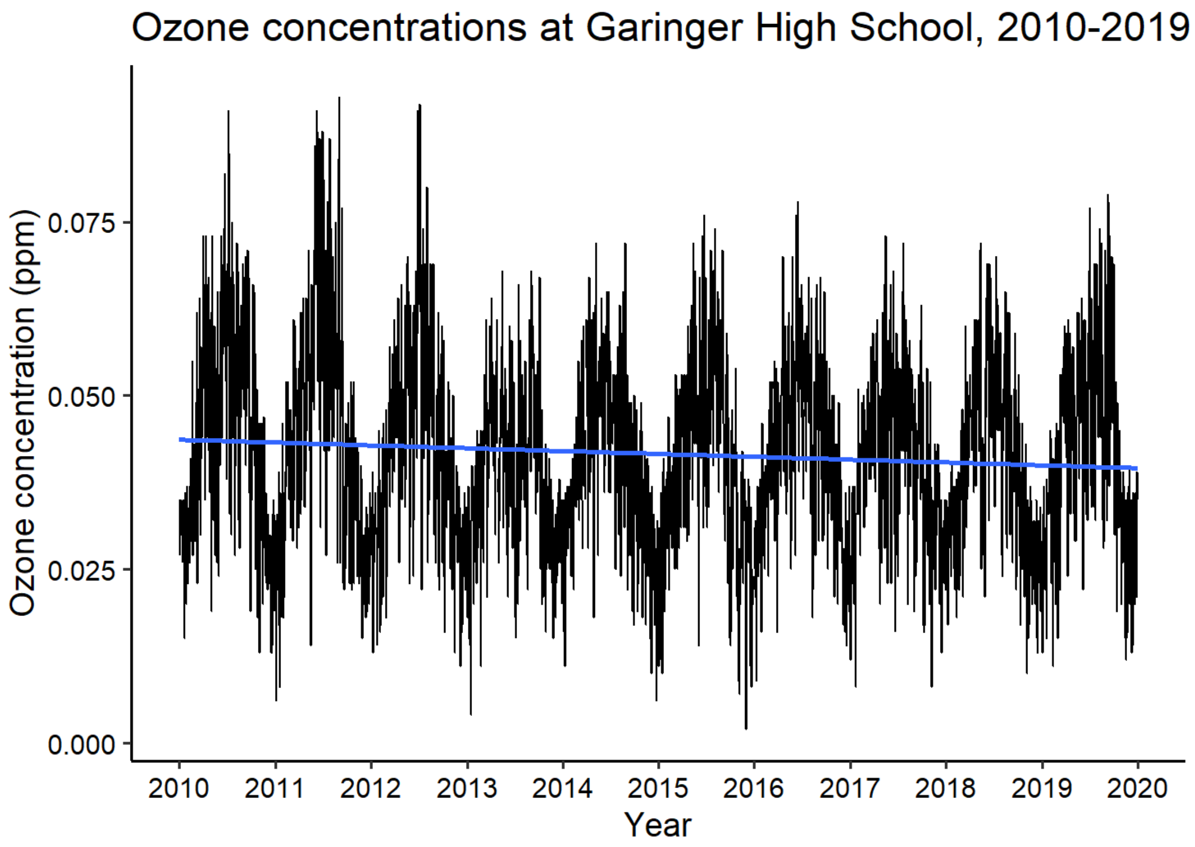
Visualize

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
ggplot(GaringerOzone, aes(x = Date, y = Daily.Max.8.hour.Ozone.Concentration)) +
  geom_line() +
  geom_smooth(method = "lm", se = FALSE) +
  scale_x_date(date_breaks = "1 year", date_labels = "%Y") +
  labs(x = "Year", y = "Ozone concentration (ppm)",
       title = "Ozone concentrations at Garinger High School, 2010-2019")
```

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

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



```
coef(summary(lm(Daily.Max.8.hour.Ozone.Concentration ~ Date, data = GaringerOzone)))[2,1]
```

```
## [1] -1.120002e-06
```

> Answer: The resulting plot shows a generally decreasing trend in ozone concentrations over time with regular periodic fluctuations from year to year. The smoothed line indicates a negative trend with a rather small slope (-1.12e-06) in ozone concentrations over the 2010-2019 time period.

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

> Answer: Linear interpolation is a simple and effective way to estimate missing values in a time series. The method is appropriate because ozone concentrations vary smoothly over time.
>
> Piecewise constant interpolation would not be appropriate because ozone concentrations can vary smoothly over time, and a piecewise constant function would not capture this behavior.
>
> Spline interpolation can be a good choice for smoothing data, but it may not be appropriate for filling in missing values because it can introduce spurious fluctuations in the data.
>
> Thus, linear interpolation strikes a good balance between simplicity and effectiveness and is often used in practice for filling in missing data.

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 %>%
  mutate(year = year(Date), month = month(Date)) %>%
  group_by(year, month) %>%
  summarise(mean_ozone = mean(Ozone.clean))
```

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

```
GaringerOzone.monthly$Date <- as.Date(paste0(format(as.numeric(GaringerOzone.monthly$year)), "-", format
(as.numeric(GaringerOzone.monthly$month)), "-01"))

head(GaringerOzone.monthly)
```

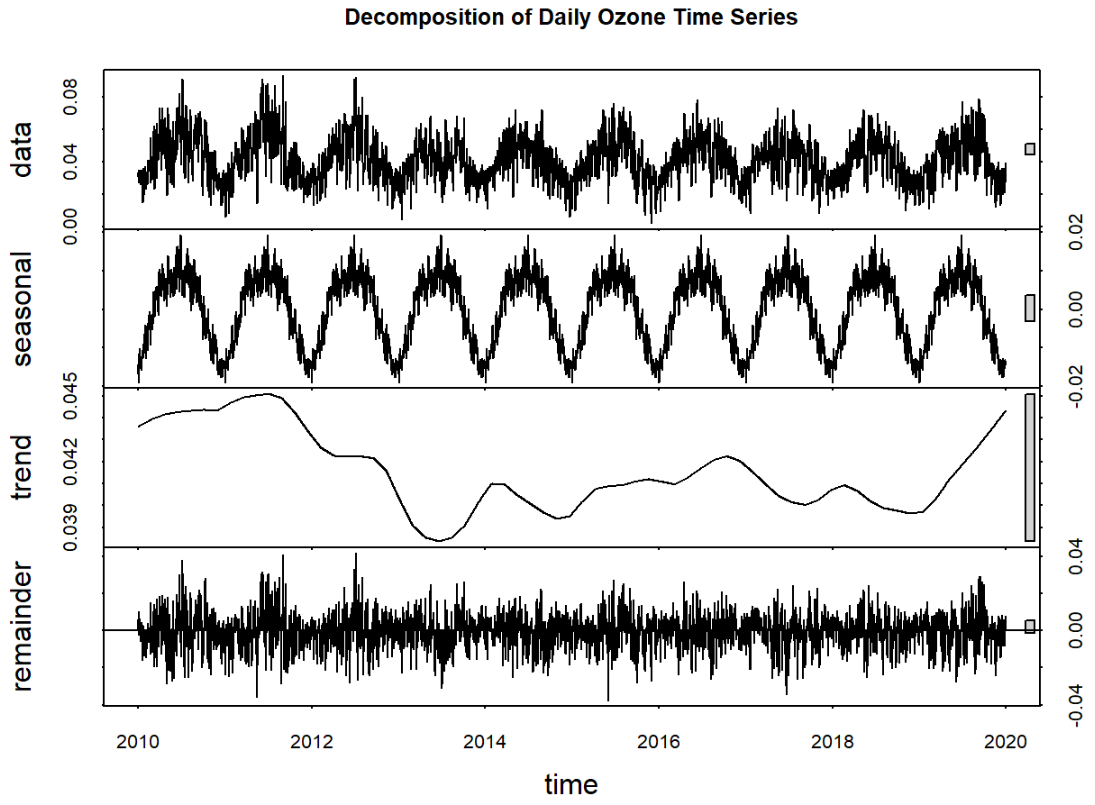
```
## # A tibble: 6 × 4
## # Groups:   year [1]
##    year month mean_ozone Date
##   <dbl> <dbl>      <dbl> <date>
## 1  2010     1     0.0305 2010-01-01
## 2  2010     2     0.0345 2010-02-01
## 3  2010     3     0.0446 2010-03-01
## 4  2010     4     0.0556 2010-04-01
## 5  2010     5     0.0466 2010-05-01
## 6  2010     6     0.0576 2010-06-01
```

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.

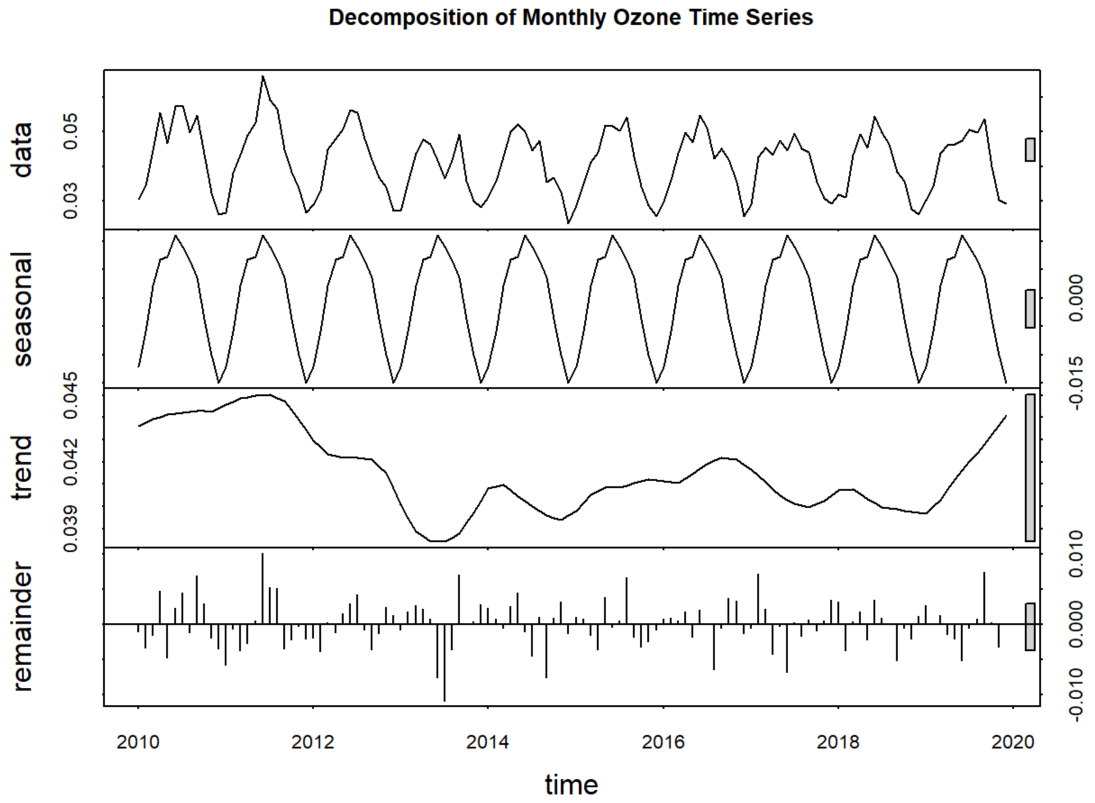
```
#10
drop_dates <- as.Date(c("2012-02-29", "2016-02-29"))
GaringerOzone <- GaringerOzone[!(GaringerOzone$Date %in% drop_dates), ]

GaringerOzone.daily.ts <- ts(GaringerOzone$Ozone.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.

```
#11
GaringerOzone.daily.decomp <- stl(GaringerOzone.daily.ts, s.window = "periodic")
plot(GaringerOzone.daily.decomp, main="Decomposition of Daily Ozone Time Series")
```



```
GaringerOzone.monthly.decomp <- stl(GaringerOzone.monthly.ts, s.window = "periodic")
plot(GaringerOzone.monthly.decomp, main="Decomposition of Monthly Ozone Time Series")
```



12. Run a monotonic trend analysis for the monthly Ozone series. In this case the seasonal Mann-Kendall is most appropriate; why is this?

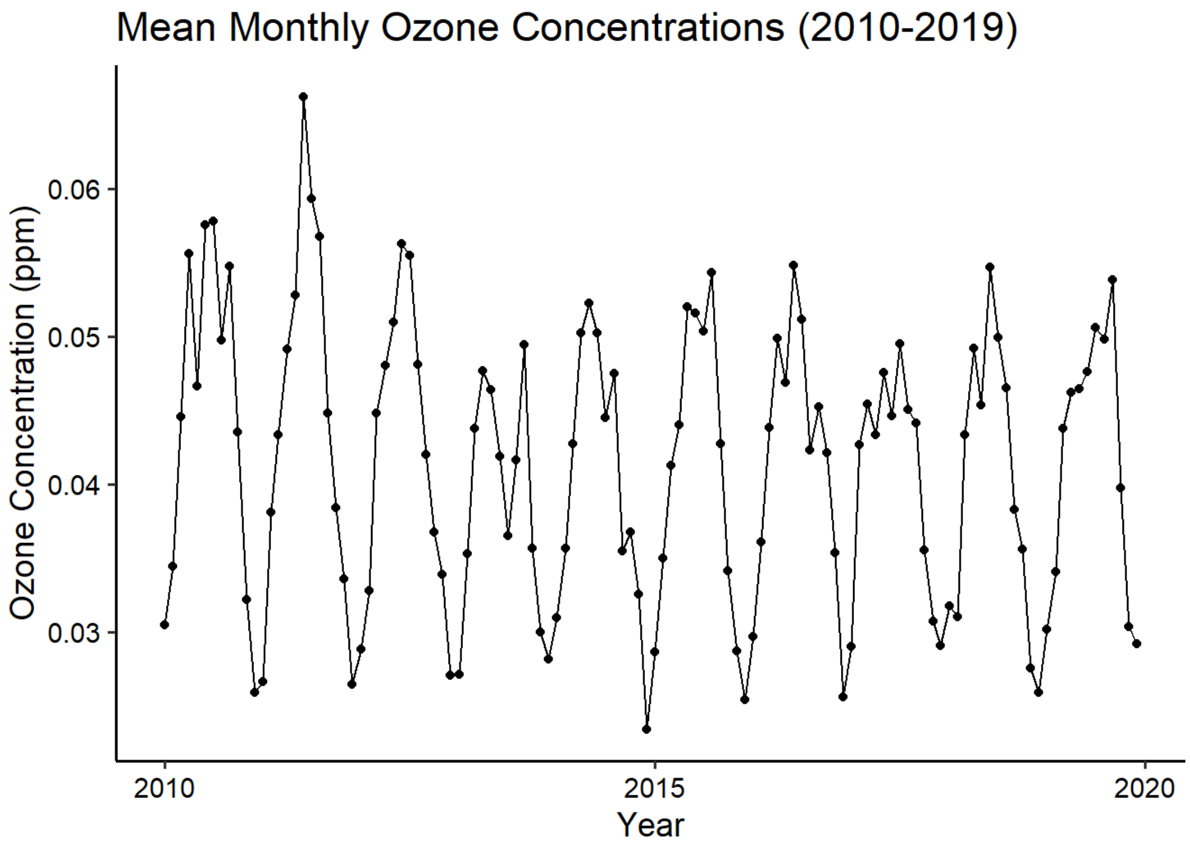
```
#12
monthly.ozone.trend <- Kendall::SeasonalMannKendall(GaringerOzone.monthly.ts)

summary(monthly.ozone.trend)
```

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

> Answer: The Mann-Kendall test is a non-parametric test used to detect trends in time series data, particularly when the data is not normally distributed or has missing values. It tests whether there is a monotonic trend (i.e., a trend in one direction) over time. **The seasonal Mann-Kendall test extends the standard Mann-Kendall test by taking into account seasonal variations in the data. This is particularly important in the case of environmental data, such as ozone concentrations, which often exhibit strong seasonal patterns.**
>
> In contrast, linear regression assumes that the relationship between the independent and dependent variables is linear and that the residuals are normally distributed. It does not take into account seasonality and is not appropriate for non-linear relationships or non-normal data.
>
> Spearman’s rank correlation coefficient is a non-parametric test that measures the strength and direction of the association between two variables. Like the Mann-Kendall test, it does not assume a linear relationship between the variables and can handle non-normal data and missing values. However, it does not take into account seasonality in the data.
>
> In summary, the choice of method for trend analysis depends on the specific characteristics of the data and the research question. If the data exhibits strong seasonality (like this case), the seasonal Mann-Kendall test is most appropriate. If the data is not normally distributed or has missing values, non-parametric tests such as the Mann-Kendall test or Spearman’s rank correlation coefficient may be more appropriate. If the data has a linear relationship and meets the assumptions of a parametric test, linear regression may be appropriate.

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 = mean_ozone)) +
  geom_point() +
  geom_line() +
  labs(x = "Year", y = "Ozone Concentration (ppm)", title = "Mean Monthly Ozone Concentrations (2010-2019)")
```



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: The graph of mean monthly ozone concentrations over time indicates that ozone concentrations at the Garinger station have been generally decreasing over the 2010s. The seasonal Mann-Kendall trend analysis confirms this trend, showing a statistically significant decreasing trend over the 10-year period (tau = -0.143, p = 0.046724 < 0.05).

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
GaringerOzone.monthly.seasonal <- as.numeric(GaringerOzone.monthly.decomp$time.series[, "seasonal"])

# Subtract the seasonal component from the original time series
GaringerOzone.monthly.nonseasonal <- as.numeric(GaringerOzone.monthly.ts - GaringerOzone.monthly.seasona
l)

#16
GaringerOzone.monthly.mk <- MannKendall(GaringerOzone.monthly.nonseasonal)
summary(GaringerOzone.monthly.mk)
```

```
## 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 yields a higher absolute score of -1179 and a lower denominator of 7139.5 compared to the seasonal Mann Kendall test on the complete series, which had a score of -77 and a denominator of 539.4972. The tau values are -0.165 and -0.143, respectively, indicating a stronger negative trend in the non-seasonal series. Furthermore, the p-value for the non-seasonal series is 0.0075402, which is smaller than the p-value of 0.046724 for the seasonal series. This suggests that the non-seasonal series exhibits a statistically significant decreasing trend over time, while the seasonal series only shows a marginally significant trend. Therefore, it may be more appropriate to use the non-seasonal Mann Kendall test for further analysis of trends in the ozone concentrations.