

Assignment 7: Time Series Analysis

Ariel Lam

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 Tuesday, March 16 at 11:59 pm.

Set up

1. Set up your session:
 - Check your working directory
 - Load the tidyverse, lubridate, zoo, and trend packages
 - Set your ggplot theme
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
getwd()

## [1] "/Users/ariellam/Desktop/EDA-Fall2022/Assignments"

setwd("~/Desktop/EDA-Fall2022")
library(tidyverse)
library(lubridate)
library(zoo)
library(trend)
library(Kendall)
library(tseries)

assignment_theme <- theme_light(base_size = 12) +
  theme(axis.text = element_text(color = "black"))

theme_set(assignment_theme)

# 2
```

```

ozone_1 <- read.csv("../Data/Raw/Ozone_TimeSeries/EPAair_03_GaringerNC2010_raw.csv")
ozone_2 <- read.csv("../Data/Raw/Ozone_TimeSeries/EPAair_03_GaringerNC2011_raw.csv")
ozone_3 <- read.csv("../Data/Raw/Ozone_TimeSeries/EPAair_03_GaringerNC2012_raw.csv")
ozone_4 <- read.csv("../Data/Raw/Ozone_TimeSeries/EPAair_03_GaringerNC2013_raw.csv")
ozone_5 <- read.csv("../Data/Raw/Ozone_TimeSeries/EPAair_03_GaringerNC2014_raw.csv")
ozone_6 <- read.csv("../Data/Raw/Ozone_TimeSeries/EPAair_03_GaringerNC2015_raw.csv")
ozone_7 <- read.csv("../Data/Raw/Ozone_TimeSeries/EPAair_03_GaringerNC2016_raw.csv")
ozone_8 <- read.csv("../Data/Raw/Ozone_TimeSeries/EPAair_03_GaringerNC2017_raw.csv")
ozone_9 <- read.csv("../Data/Raw/Ozone_TimeSeries/EPAair_03_GaringerNC2018_raw.csv")
ozone_10 <- read.csv("../Data/Raw/Ozone_TimeSeries/EPAair_03_GaringerNC2019_raw.csv")

GaringerOzone <- rbind(ozone_1, ozone_2, ozone_3,
  ozone_4, ozone_5, ozone_6, ozone_7, ozone_8,
  ozone_9, ozone_10)

```

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 <- seq.Date(as.Date("2010-01-01"), as.Date("2019-12-31"),
  "days")

Days <- as.data.frame(Days) %>%
  rename(Date = Days)

# 6
GaringerOzone <- Days %>%
  left_join(GaringerOzone)

```

```
## Joining, by = "Date"
```

Visualize

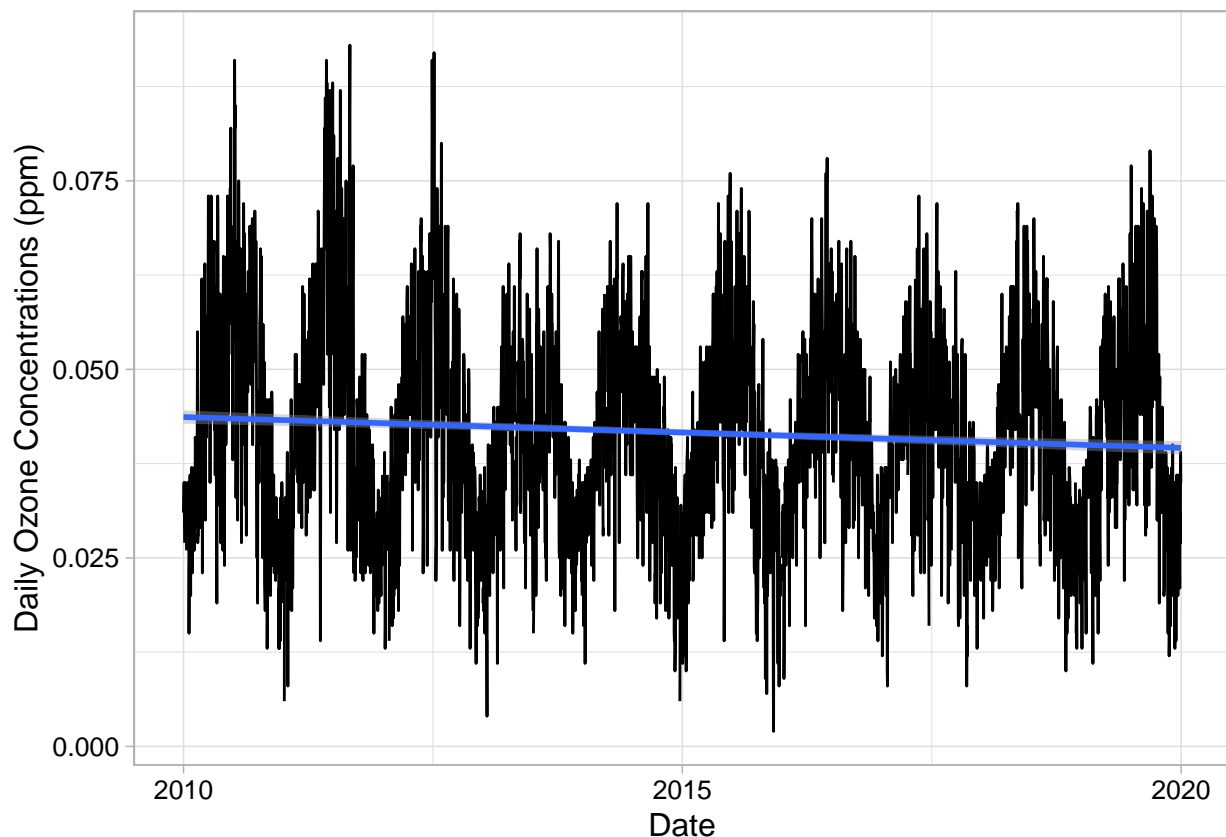
7. Create a line plot depicting ozone concentrations over time. In this case, we will plot actual concentrations in ppm, not AQI values. Format your axes accordingly. Add a smoothed line showing any linear trend

of your data. Does your plot suggest a trend in ozone concentration over time?

```
# 7
ozone_over_time <- ggplot(GaringerOzone, aes(x = Date,
  y = Daily.Max.8.hour.Ozone.Concentration)) +
  geom_line() + ylab("Daily Ozone Concentrations (ppm)") +
  geom_smooth(method = "lm")

ozone_over_time

## `geom_smooth()` using formula 'y ~ x'
## Warning: Removed 63 rows containing non-finite values (stat_smooth).
```



Answer: There is a slight negative trend detected with the linear smooth line. This suggests that daily ozone values might be slowly decreasing over time.

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(Daily.Max.8.hour.Ozone.Concentration = zoo::na.approx(Daily.Max.8.hour.Ozone.Concentration))
```

Answer: A piecewise uses the “nearest neighbor approach” and takes the data from either earlier or later and replaces the missing data with the value. Spline uses a quadratic function to interpolate

data. The piecewise wasn't used due to the fact that the concentrations fluctuate daily, so a nearest neighbor approach may not be appropriate. The graph shows a linear trend, and therefore should not be interpolated quadratically or the values would not be a good estimate.

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(Month = month(Date)) %>%
  mutate(Year = year(Date)) %>%
  group_by(Year, Month) %>%
  summarise(Mean_Ozone = mean(Daily.Max.8.hour.Ozone.Concentration))

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

```
GaringerOzone.monthly <- GaringerOzone.monthly %>%
  mutate(Month_Year = make_date(month = Month,
    year = Year))
```

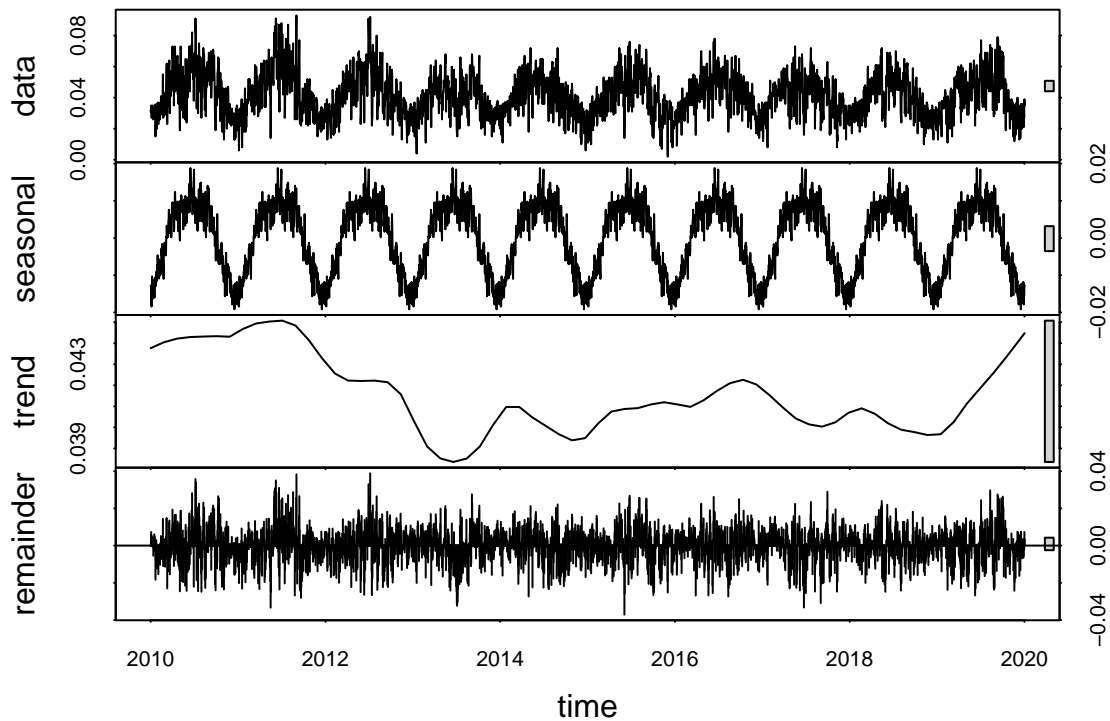
10. Generate two time series objects. Name the first `GaringerOzone.daily.ts` and base it on the dataframe of daily observations. Name the second `GaringerOzone.monthly.ts` and base it on the monthly average ozone values. Be sure that each specifies the correct start and end dates and the frequency of the time series.

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

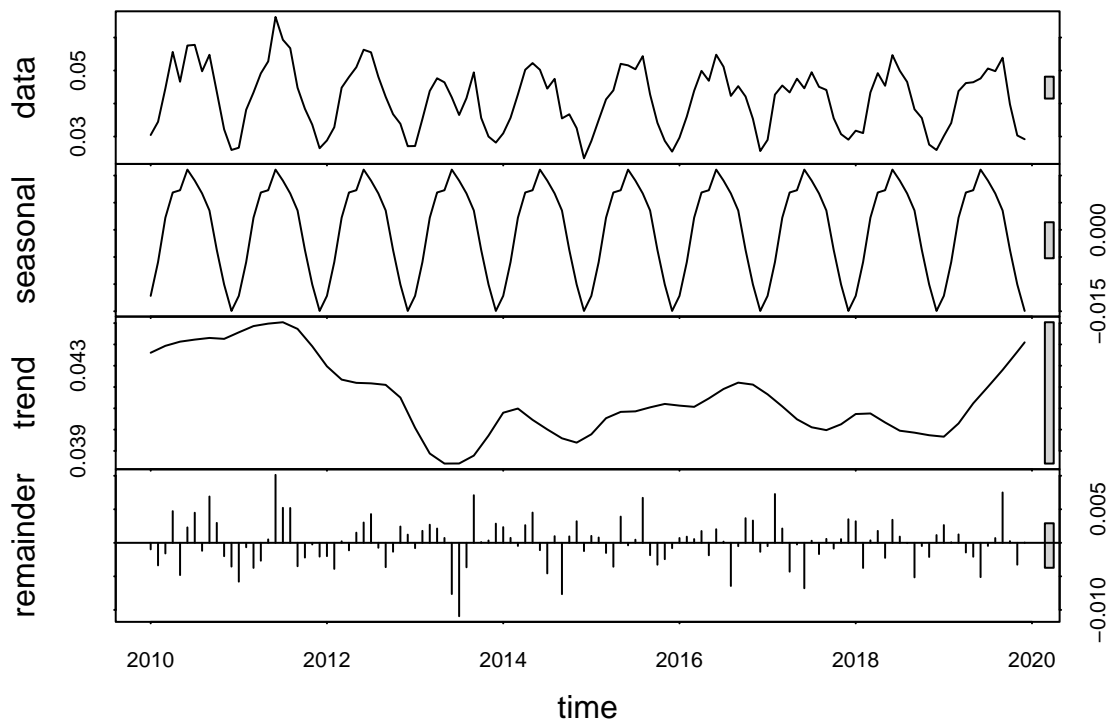
GaringerOzone.monthly.ts <- ts(GaringerOzone.monthly$Mean_Ozone,
  start = c(2010, 1), frequency = 12)
```

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

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



```
Garinger_Ozone_Monthly_decomp <- stl(GaringerOzone.monthly.ts,
  s.window = "periodic")
plot(Garinger_Ozone_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
monthly_ozone_monotonic_trend_analysis <- Kendall::SeasonalMannKendall(GaringerOzone.monthly.ts)
```

```
monthly_ozone_monotonic_trend_analysis
```

```
## tau = -0.143, 2-sided pvalue =0.046724
```

```
summary(monthly_ozone_monotonic_trend_analysis)
```

```
## Score = -77 , Var(Score) = 1499
```

```
## denominator = 539.4972
```

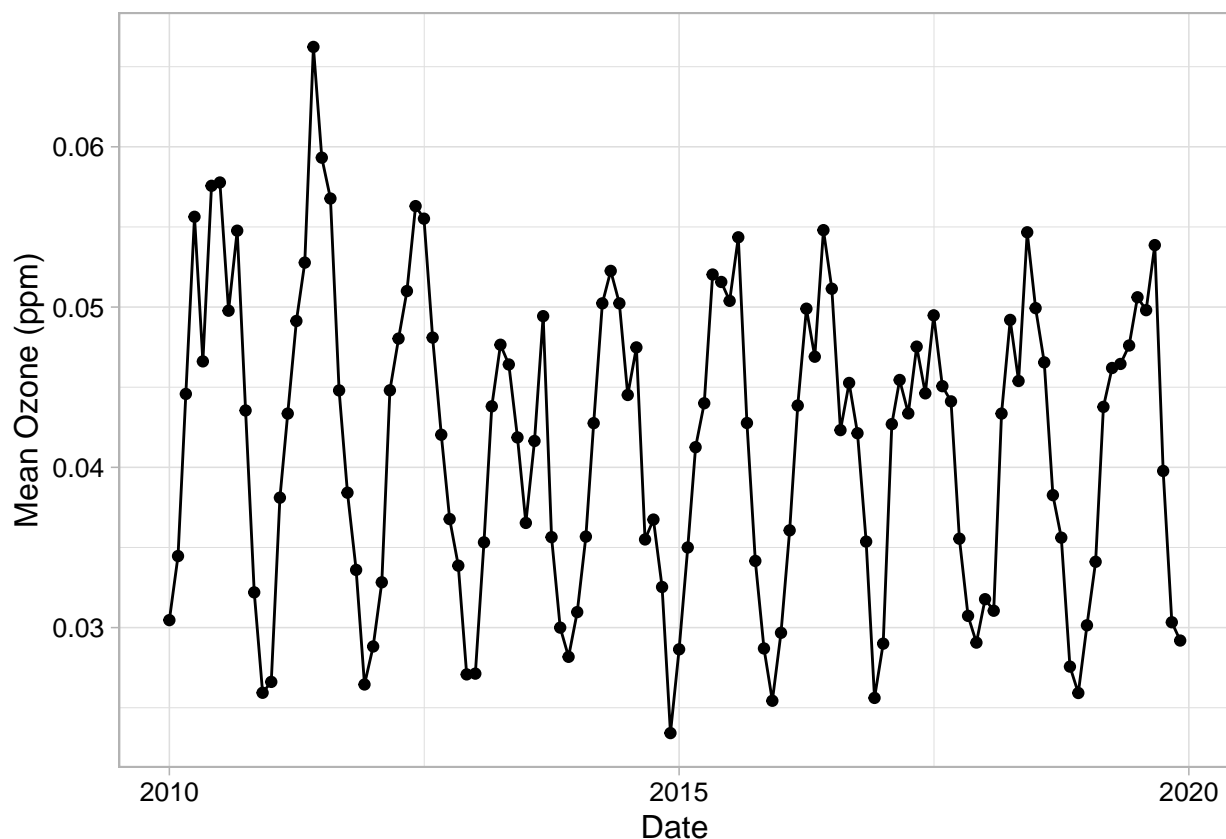
```
## tau = -0.143, 2-sided pvalue =0.046724
```

Answer: From the time series plots it appears that seasonal monthly patterns occur at regular intervals, so a seasonal Mann-Kendall trend would show a clear trend by comparing data between specific seasons. The Mann-Kendall will show if a monotonic trend exists. Seasonal Mann-Kendall is also non-parametric so does not assume anything about the dataset distribution.

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(y = Mean_Ozone,  
  x = Month_Year)) + geom_point() + geom_line() +  
  ylab("Mean Ozone (ppm)") + xlab("Date")
```



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 research inquiry of whether ozone concentrations have decreased over time at this specific station and the significant p-score value and the negative score indicate that mean ozone

is slowly decreasing over time when seasonality is taken into account. ($\tau = -0.143$, 2-sided $pvalue = 0.046724$, $Score = -77$, $Var(Score) = 1499$, denominator = 539.4972 $\tau = -0.143$, 2-sided $pvalue = 0.046724$)

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_Ozone_Monthly_decomp$time.series[,  
  2:3])  
Garinger_Components <- mutate(Garinger_Components,  
  Observed = GaringerOzone.monthly$Mean_Ozone,  
  Date = GaringerOzone.monthly$Month_Year)  
GaringerOzone.nonseasonal.ts <- ts(Garinger_Components$Observed,  
  start = c(2010, 1), frequency = 12)
```

16

```
non_seasonal_ozone_components_analysis <- MannKendall(GaringerOzone.nonseasonal.ts)  
non_seasonal_ozone_components_analysis
```

```
## tau = -0.0594, 2-sided pvalue = 0.33732
```

```
summary(non_seasonal_ozone_components_analysis)
```

```
## Score = -424 , Var(Score) = 194364.7
```

```
## denominator = 7139
```

```
## tau = -0.0594, 2-sided pvalue = 0.33732
```

Answer: The non-seasonal ozone monthly series has a p-value that is not significant. Therefore the null hypothesis of ozone amount not changing must be accepted. This shows that seasonality influences the trend of decreasing ozone and you need to look at the trends between the same seasons and not just the data overall. This could be due to the fact that weather or precipitation may impact the presence of ozone.