Assignment 10: Data Scraping

Rachael Stephan

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on data scraping.

Directions

1. Rename this file `<FirstLast>_A10_DataScraping.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 your code is tidy; use line breaks to ensure your code fits in the knitted output.
5. Be sure to **answer the questions** in this assignment document.
6. When you have completed the assignment, **Knit** the text and code into a single PDF file.

Set up

1. Set up your session:

- Load the packages `tidyverse`, `rvest`, and any others you end up using.
- Check your working directory

```
#1
#load libraries
library(tidyverse); library(rvest); library(here); library(lubridate)

#check working directory
here()
```

```
## [1] "/home/guest/Duke_R/EDE_Fall2024"
```

```
getwd()
```

```
## [1] "/home/guest/Duke_R/EDE_Fall2024"
```

```
#set chunk options
knitr::opts_chunk$set(warning = FALSE, message = FALSE)

#set plot theme
mytheme <- theme_bw(base_size = 10)+
  theme(axis.title = element_text(size = 10, hjust = 0.5),
        plot.title.position = "panel",
        plot.caption = element_text(hjust = 0),
        legend.box = "vertical",
        legend.location = "plot",
        axis.gridlines = element_line(color = "grey", linewidth = 0.25),
        axis.ticks = element_line(color = "black", linewidth = 0.5),
        axis.text.x = element_text(angle = 45, hjust = 1))
theme_set(mytheme)
```

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham’s 2023 Municipal Local Water Supply Plan (LWSP):

- Navigate to https://www.ncwater.org/WUDC/app/LWSP/search.php
- Scroll down and select the LWSP link next to Durham Municipality.
- Note the web address: https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2023

Indicate this website as the as the URL to be scraped. (In other words, read the contents into an `rvest` webpage object.)

```
#2
#set the webpage
webpage <- read_html("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2023")
```

3. The data we want to collect are listed below:

- From the “1. System Information” section:

- Water system name

- PWSID

- Ownership

- From the “3. Water Supply Sources” section:

- Maximum Day Use (MGD) - for each month

In the code chunk below scrape these values, assigning them to four separate variables.

> HINT: The first value should be “Durham”, the second “03-32-010”, the third “Municipality”, and the last should be a vector of 12 numeric values (represented as strings)“.

```
#3
#scrape the water system data
water_system <- webpage %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()

#scrape the PWSID
pwsid <- webpage %>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
  html_text()
```

```
#scrape the system ownership
system_ownership <- webpage %>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%
  html_text()

#scrape the max daily use for each month
mdu.month <- webpage %>%
  html_nodes("th~ td+ td") %>%
  html_text()
```

4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

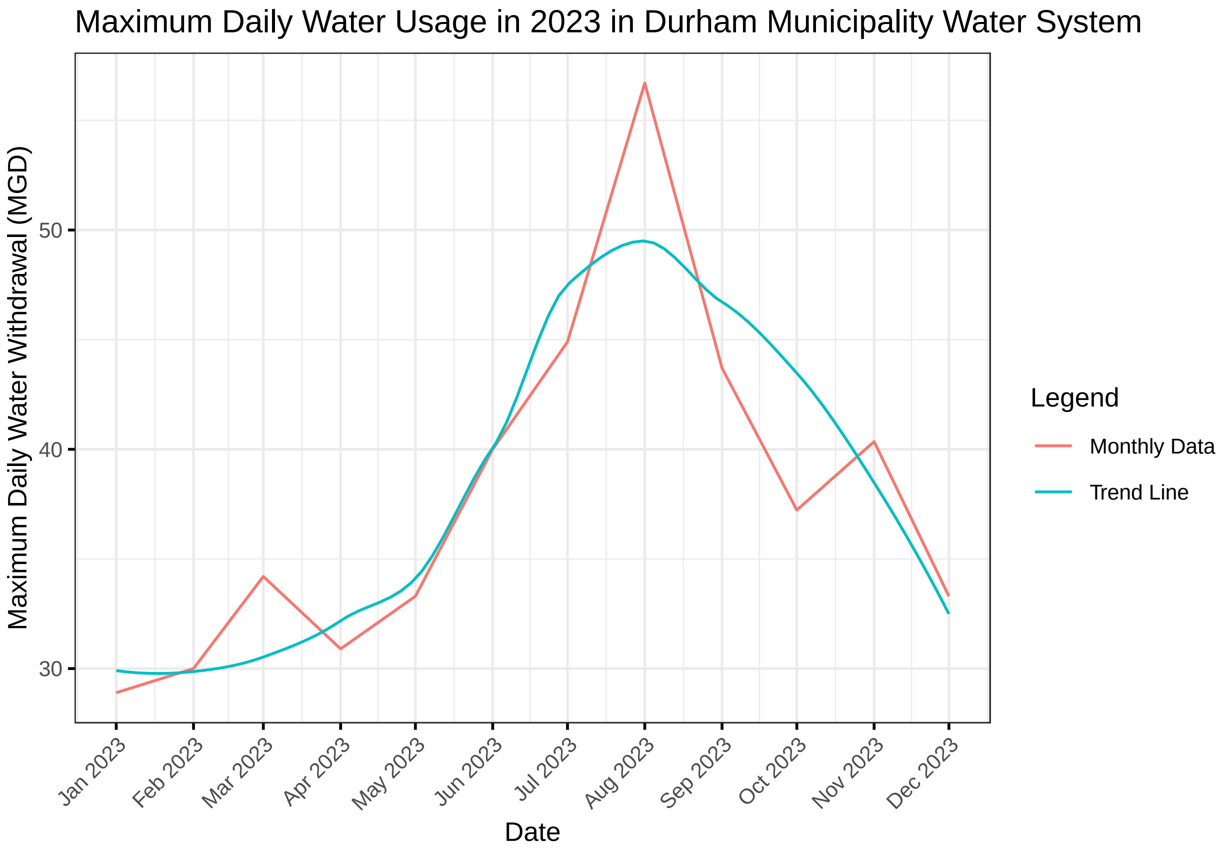
   TIP: Use `rep()` to repeat a value when creating a dataframe.

   NOTE: It's likely you won't be able to scrape the monthly widthrawal data in chronological order. You can overcome this by creating a month column manually assigning values in the order the data are scraped: "Jan", "May", "Sept", "Feb", etc… Or, you could scrape month values from the web page…

5. Create a line plot of the maximum daily withdrawals across the months for 2023, making sure, the months are presented in proper sequence.

```
#4
#create month vector to assign to data frame
mon <- webpage %>%
  html_nodes(".fancy-table:nth-child(31) tr+ tr th") %>%
  html_text()

#create data frame. Set the date to the first of each month for easy plotting
max_water_use <- data.frame("Month" = mon,
                            "Year" =  rep(2023, 12),
                            "Date" = my(paste(mon, "2023")),
                            "Water.System" = rep(water_system,12),
                            "PWSID" = rep(pwsid, 12),
                            "System.Ownership" = rep(system_ownership, 12),
                            "Max_Withdrawals_mgd" = as.numeric(mdu.month))

#5
#create line plot
max_use.plot <- ggplot(data = max_water_use, aes(x= Date, y = Max_Withdrawals_mgd))+
  geom_line(aes(colour = "Monthly Data"))+
  geom_smooth(aes(colour = "Trend Line"),se = F, size = 0.5)+
  labs(title = "Maximum Daily Water Usage in 2023 in Durham Municipality Water System",
       y = "Maximum Daily Water Withdrawal (MGD)",
       colour = "Legend")+
  scale_x_date(breaks = seq(from = min(max_water_use$Date),
                            to = max(max_water_use$Date),
                            by = "1 month"),
               date_labels = "%b %Y")
```

```
max_use.plot
```

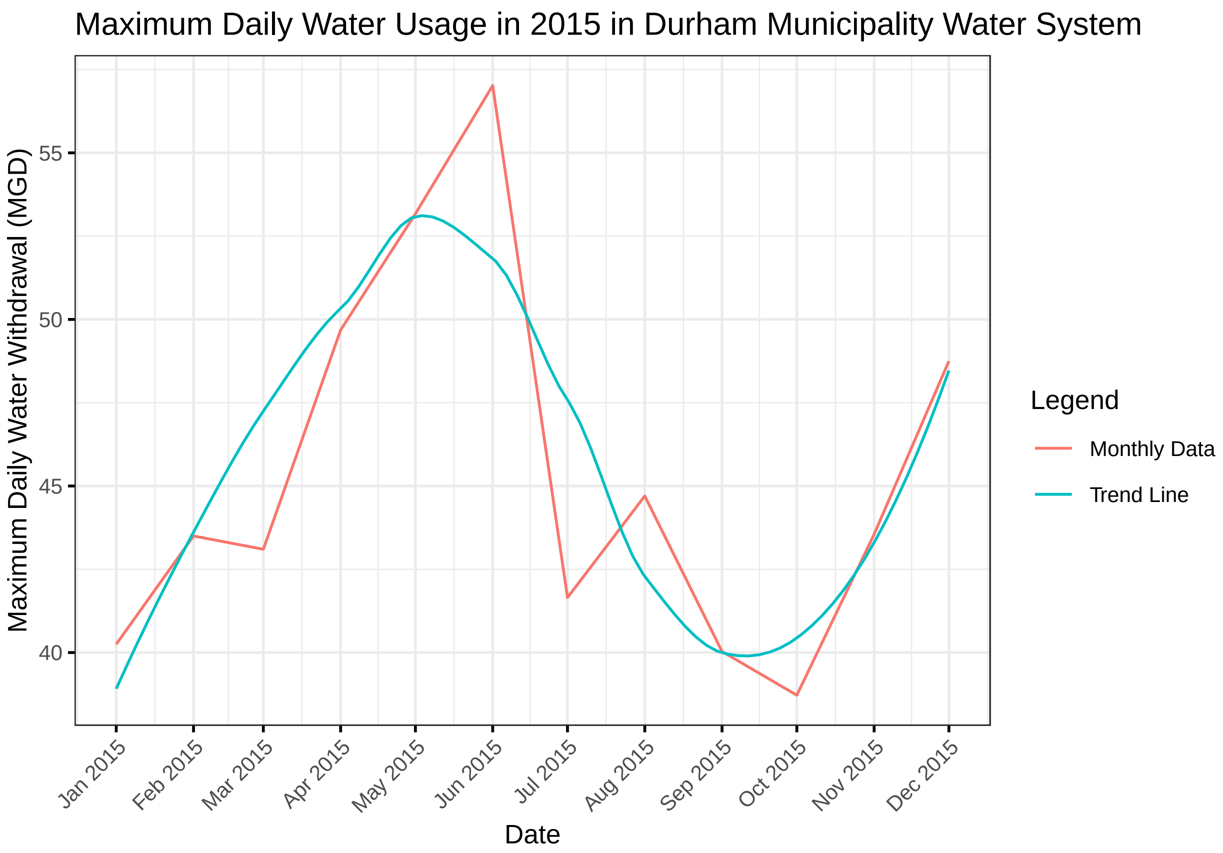


6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data, returning a dataframe. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

```
#6.
#create fuction
PWSID.year.scrape <- function(my_pwsid, my_year) {
  #create url and read it
  base_url <- "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid="
  scrape_url <- paste0(base_url, my_pwsid, "&year=", my_year)
  my_webpage <- read_html(scrape_url)

  #get the relevant tag info
  #water system data
  my_water_system <- my_webpage %>%
    html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
    html_text()

  #PWSID is provide and does not need to be scraped

  #system ownership
  my_system_ownership <- my_webpage %>%
   html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%
   html_text()

  #scrape the max daily use for each month
  my_mdu.month <- my_webpage %>%
    html_nodes("th~ td+ td") %>%
    html_text()

  #create month vector manually due to difficulties with scraping as per John
  my_mon <- c("Jan", "May", "Sep", "Feb", "Jun", "Oct",
              "Mar", "Jul", "Nov", "Apr", "Aug", "Dec")

  #create data frame. Set the date to the first of each month for easy plotting
  my_max_water_use <- data.frame("Month" = my_mon,
                            "Year" =  rep(my_year, 12),
                            "Date" = my(paste(my_mon, my_year)),
                            "Water.System" = rep(my_water_system,12),
                            "PWSID" = rep(my_pwsid, 12),
                            "System.Ownership" = rep(my_system_ownership, 12),
                            "Max_Withdrawals_mgd" = as.numeric(my_mdu.month))

  return(my_max_water_use)
}
```

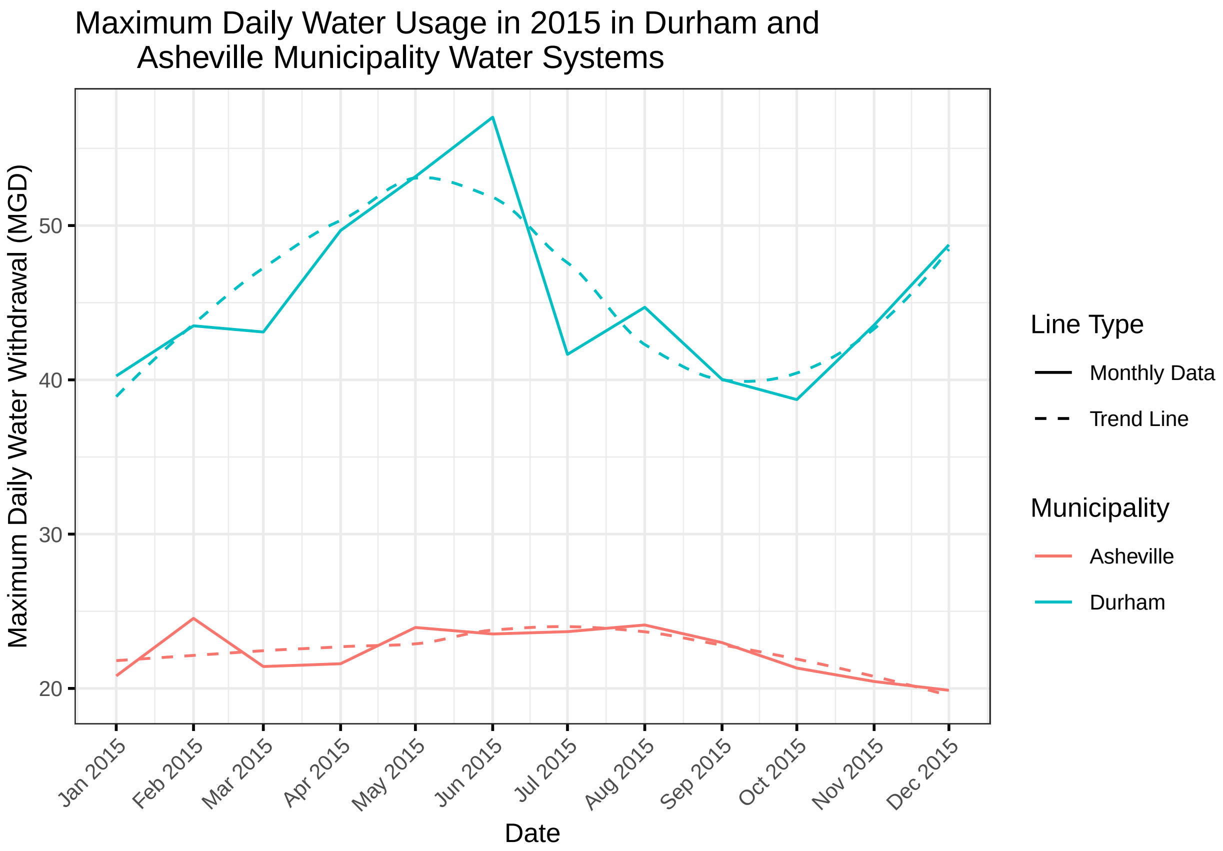
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID='03-32-010') for each month in 2015

```
#7
#run function with given PWSID and year
max_water_use.durham.2015 <- PWSID.year.scrape("03-32-010", 2015)

#create function to automatically plot scraped data
durham.2015.plot <- ggplot(data = max_water_use.durham.2015,
                           aes(x= Date, y = Max_Withdrawals_mgd))+
  geom_line(aes(colour = "Monthly Data"))+
  geom_smooth(aes(colour = "Trend Line"), se = F, size = 0.5)+
  labs(title = "Maximum Daily Water Usage in 2015 in Durham Municipality Water System",
       y = "Maximum Daily Water Withdrawal (MGD)",
       colour = "Legend")+
  scale_x_date(breaks = seq(from = min(max_water_use.durham.2015$Date),
                            to = max(max_water_use.durham.2015$Date),
                            by = "1 month"),
               date_labels = "%b %Y")

durham.2015.plot
```



8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares Asheville's to Durham's water withdrawals.

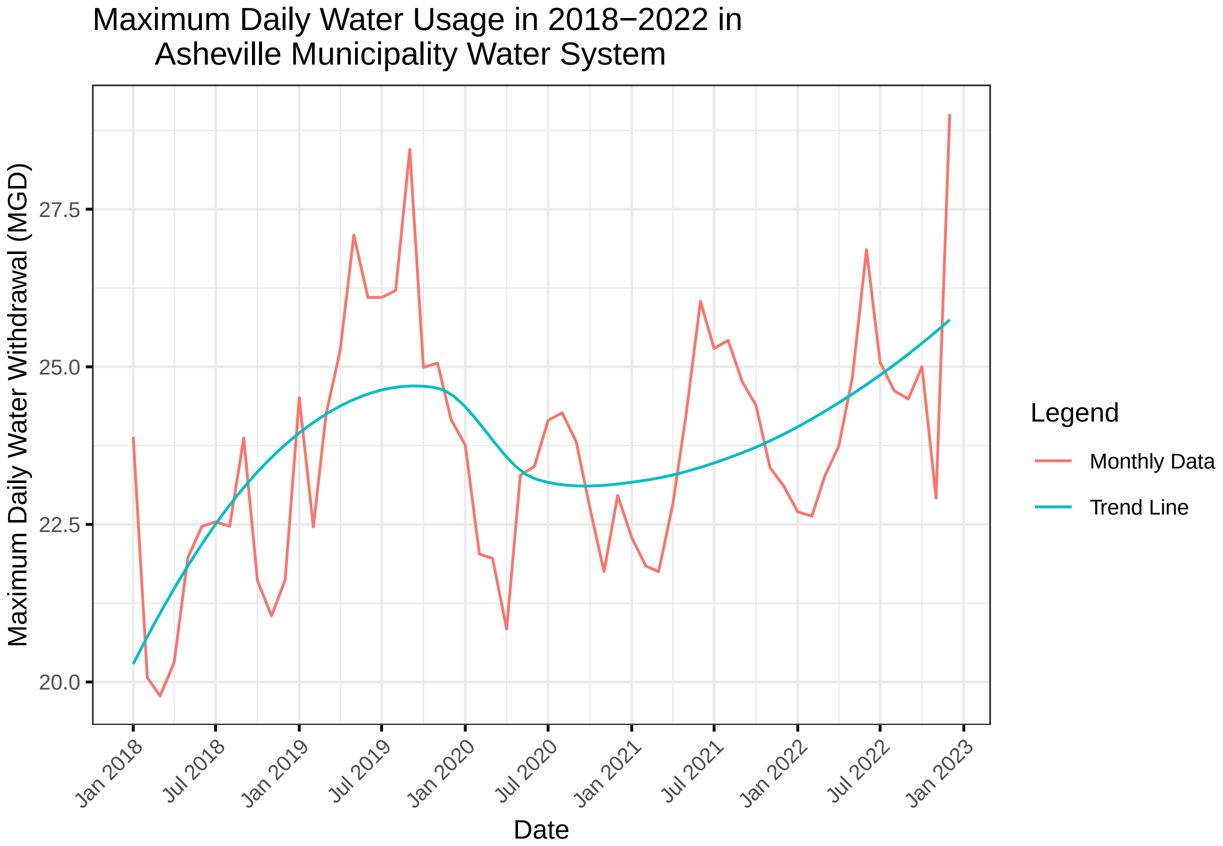
```
#8
#make a data frame about asheville
max_water_use.asheville.2015 <- PWSID.year.scrape("01-11-010", 2015)

#bind data frames by rows
max_water.durham_asheville.2015 <- rbind(max_water_use.asheville.2015,
                                         max_water_use.durham.2015)

#plot the data
durham_asheville.2015.plot <- ggplot(data = max_water.durham_asheville.2015,
                                     aes(x= Date, y = Max_Withdrawals_mgd,
                                         colour = Water.System))+
  geom_line(aes(linetype = "Monthly Data"))+
  geom_smooth(aes(linetype = "Trend Line"), se = F, size = 0.5)+
  labs(title = "Maximum Daily Water Usage in 2015 in Durham and
       Asheville Municipality Water Systems",
       y = "Maximum Daily Water Withdrawal (MGD)",
       colour = "Municipality",
       linetype = "Line Type")+
  scale_x_date(breaks = seq(from = min(max_water_use.durham.2015$Date),
                            to = max(max_water_use.durham.2015$Date),
                            by = "1 month"),
               date_labels = "%b %Y")+
  scale_linetype_manual(values = c("Trend Line" = "dashed", "Monthly Data" = "solid"))+
  guides(linetype = guide_legend(override.aes = list(colour = "black")))

durham_asheville.2015.plot
```



9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2018 thru 2022.Add a smoothed line to the plot (method = 'loess').

   TIP: See Section 3.2 in the "10_Data_Scraping.Rmd" where we apply "map2()" to iteratively run a function over two inputs. Pipe the output of the map2() function to `bindrows()` to combine the dataframes into a single one.

```
#9
#create multiple year dataframe
asheville.multiyear <- map2("01-11-010", 2018:2022, PWSID.year.scrape) %>%
  bind_rows()

asheville.2018_2022.plot <- ggplot(data = asheville.multiyear,
                                   aes(x= Date, y = Max_Withdrawals_mgd))+
  geom_line(aes(colour = "Monthly Data"))+
  geom_smooth(aes(colour = "Trend Line"), se = F, size = 0.5)+
  labs(title = "Maximum Daily Water Usage in 2018-2022 in
       Asheville Municipality Water System",
       y = "Maximum Daily Water Withdrawal (MGD)",
       colour = "Legend")+
  scale_x_date(breaks = seq(from = min(asheville.multiyear$Date),
                            to = max(asheville.multiyear$Date)+31,
                            by = "6 month"),
               date_labels = "%b %Y")

asheville.2018_2022.plot
```



Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time?

There is an upwards trend in water usage in Asheville over time.