Assignment 10: Data Scraping

Alex Lopez

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

library(tidyverse);library(lubridate);library(viridis);library(here)
library(rvest)
```

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

website <-
  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

Water.System.Name <- website %>%
  html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>%
  html_text()

PWSID <- website %>%
  html_nodes('td tr:nth-child(1) td:nth-child(5)') %>%
  html_text()

Ownership <- website %>%
  html_nodes('div+ table tr:nth-child(2) td:nth-child(4)') %>%
  html_text()

MGD <- website %>%
  html_nodes('th~ td+ td') %>%
  html_text()
```

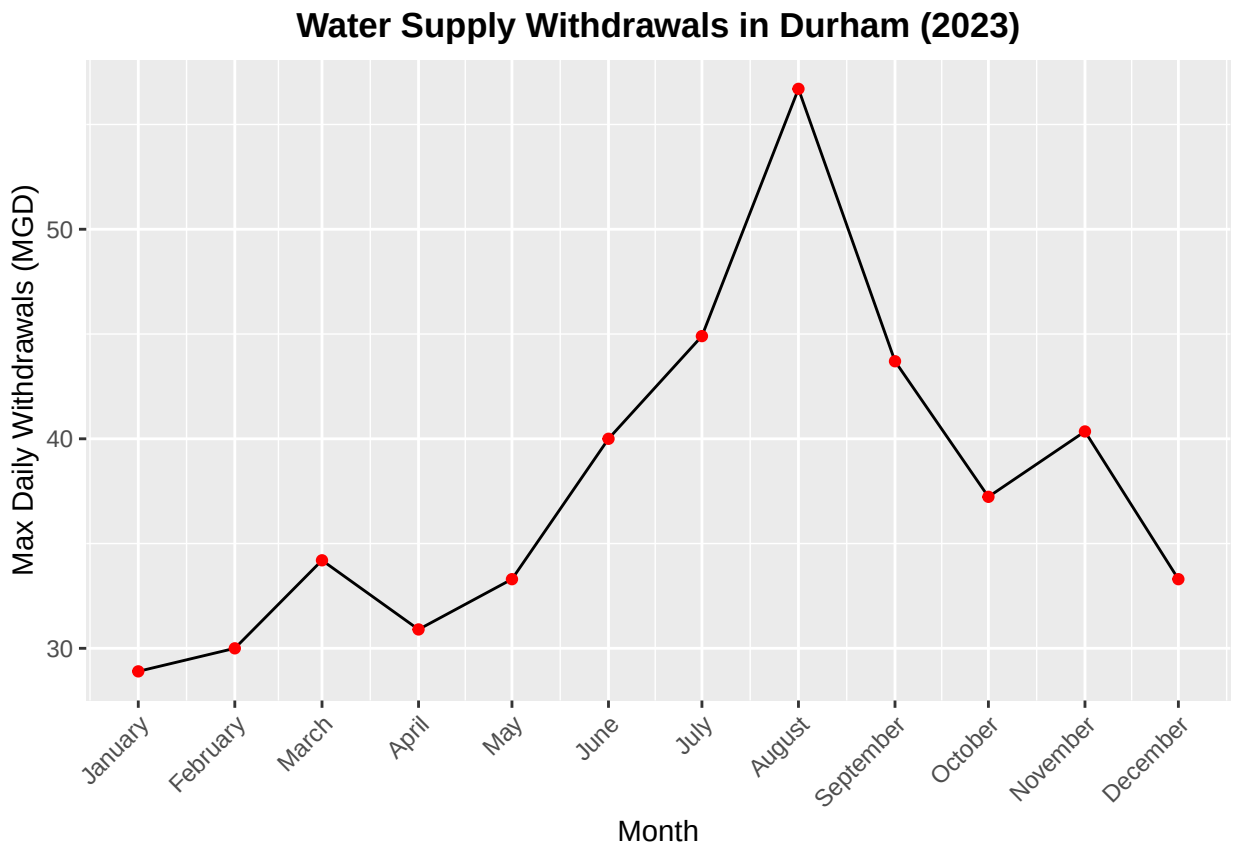
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   > 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.

```r
#4

df.scraped.data <- data.frame(Water.System.Name = rep(Water.System.Name, 12),
                              PWSID = rep(PWSID, 12),
                              Ownership = rep(Ownership, 12),
                              MGD = as.numeric(MGD),
                              Month = c("Jan", "May", "Sep", "Feb", "Jun",
                                        "Oct", "Mar", "Jul", "Nov", "Apr",
                                        "Aug", "Dec"),
                              Year = rep(2023, 12)
                              )

#add date column
df.scraped.data$Date <-
  as.Date(paste(df.scraped.data$Year, df.scraped.data$Month, '01'),
                                format = '%Y %b %d')

#5

ggplot(df.scraped.data, aes(x = Date, y = MGD)) +
  geom_line() +
  geom_point(color = "red") +
  scale_x_date(date_labels = "%B", date_breaks = "1 month") +
  labs(
    title = "Water Supply Withdrawals in Durham (2023)",
    x = "Month",
    y = "Max Daily Withdrawals (MGD)"
  ) +
  theme_gray() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title = element_text(hjust = 0.5, face = "bold")
  )
```



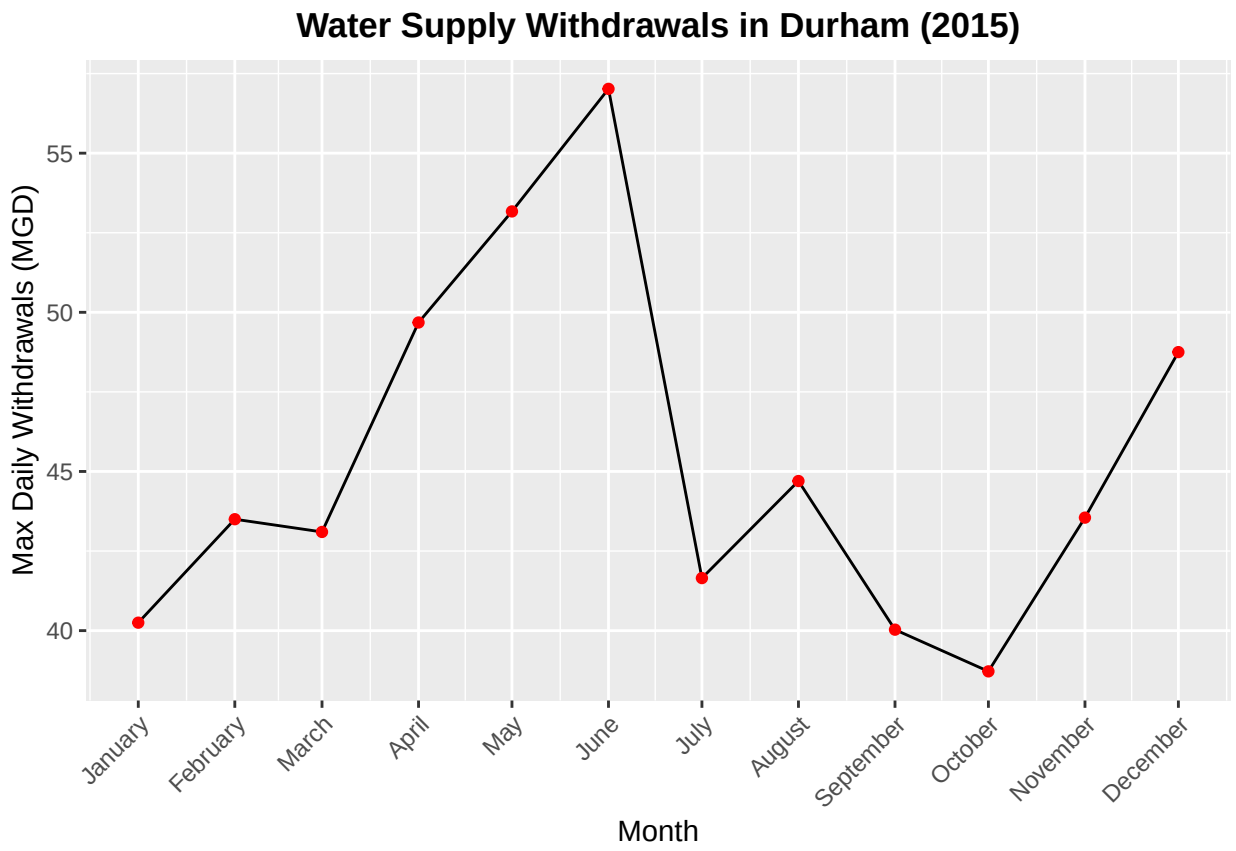
6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data, returning a dataframe. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

```
#6.

scrape.fn <- function(the.pwsid, the.year){
  the.website <-
    read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',
                                    the.pwsid, '&year=', the.year))

  #set element address variables
  water.system.name.tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'
  pwsid.tag <- 'td tr:nth-child(1) td:nth-child(5)'
  ownership.tag <- 'div+ table tr:nth-child(2) td:nth-child(4)'
  mgd.tag <- 'th~ td+ td'

  #scrape data items
  the.watersystem <-
    the.website %>% html_nodes(water.system.name.tag) %>% html_text()
  the.pwsid.number <-
    the.website %>% html_nodes(pwsid.tag) %>% html_text()
  the.ownership <-
    the.website %>% html_nodes(ownership.tag) %>% html_text()
```

```
  the.mgd.value <-
    the.website %>% html_nodes(mgd.tag) %>% html_text()

  #create df from scraped data
  df.lwsp <- data.frame("Month" = c("Jan", "May", "Sep", "Feb", "Jun", "Oct",
                                    "Mar", "Jul", "Nov", "Apr", "Aug", "Dec"),
                        "Year" = rep(the.year,12),
                        "Max Day Use" = as.numeric(the.mgd.value)) %>%
  mutate(Water.System.Name = !!the.watersystem,
         PWSID = !!the.pwsid.number,
         Ownership = !!the.ownership,
         Date = my(paste(Month,"-",Year)))

  return(df.lwsp)
}
```

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

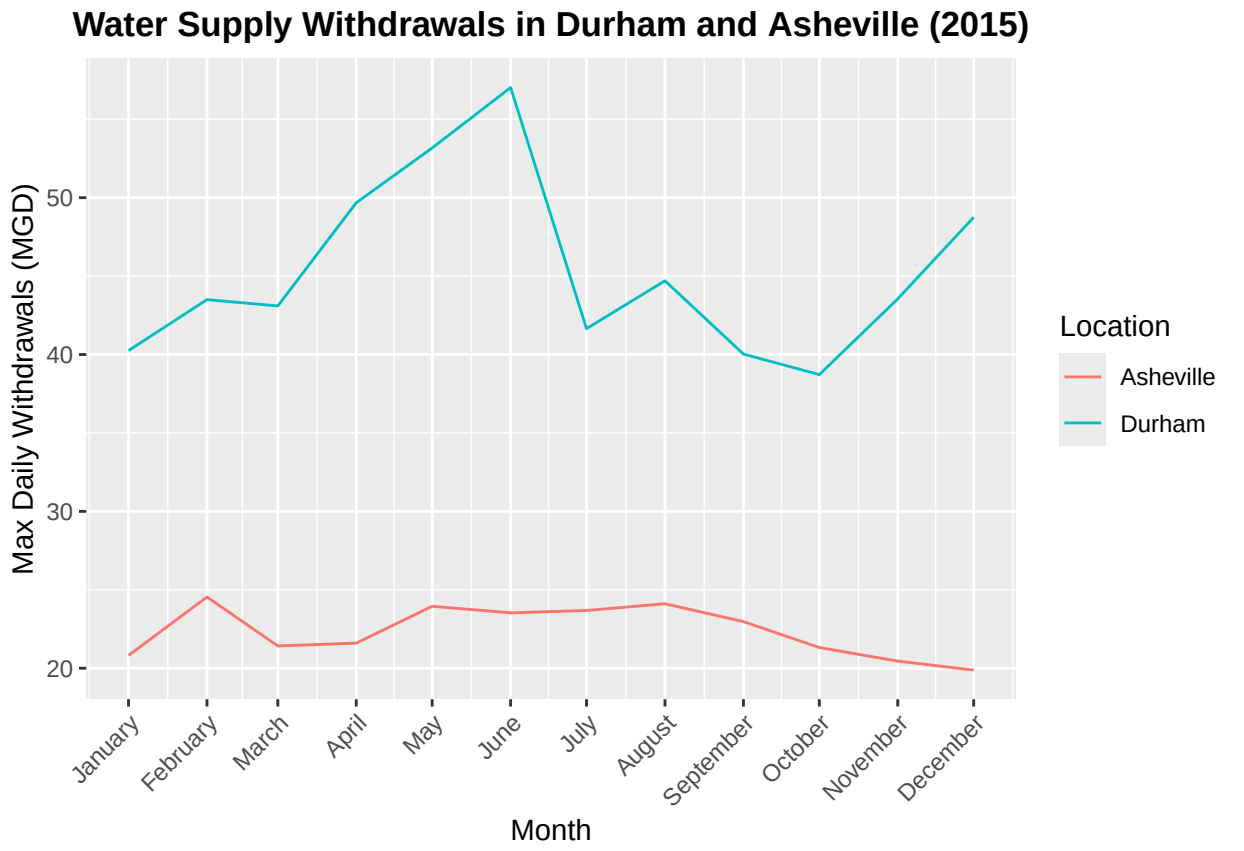
```
#7

df.lwsp.durham <- scrape.fn('03-32-010',2015)
view(df.lwsp.durham)

ggplot(df.lwsp.durham, aes(x = Date, y = Max.Day.Use)) +
  geom_line() +
  geom_point(color = "red") +
  labs(
    title = "Water Supply Withdrawals in Durham (2015)",
    x = "Month",
    y = "Max Daily Withdrawals (MGD)"
  ) +
  scale_x_date(date_labels = "%B", date_breaks = "1 month") +
  theme_gray() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title = element_text(hjust = 0.5, face = "bold")
  )
```



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

df.lwsp.asheville <- scrape.fn('01-11-010',2015)
view(df.lwsp.asheville)

#combine data for durham and asheville
df.combined <- bind_rows(
  df.lwsp.durham %>% mutate(Location = "Durham"),
  df.lwsp.asheville %>% mutate(Location = "Asheville")
)

ggplot(df.combined, aes(x = Date, y = Max.Day.Use, color = Location)) +
  geom_line() +
  labs(
    title = "Water Supply Withdrawals in Durham and Asheville (2015)",
    x = "Month",
    y = "Max Daily Withdrawals (MGD)",
    color = "Location"
  ) +
  scale_x_date(date_labels = "%B", date_breaks = "1 month") +
  theme_gray() +
```

```
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title = element_text(hjust = 0.5, face = "bold")
  )
```



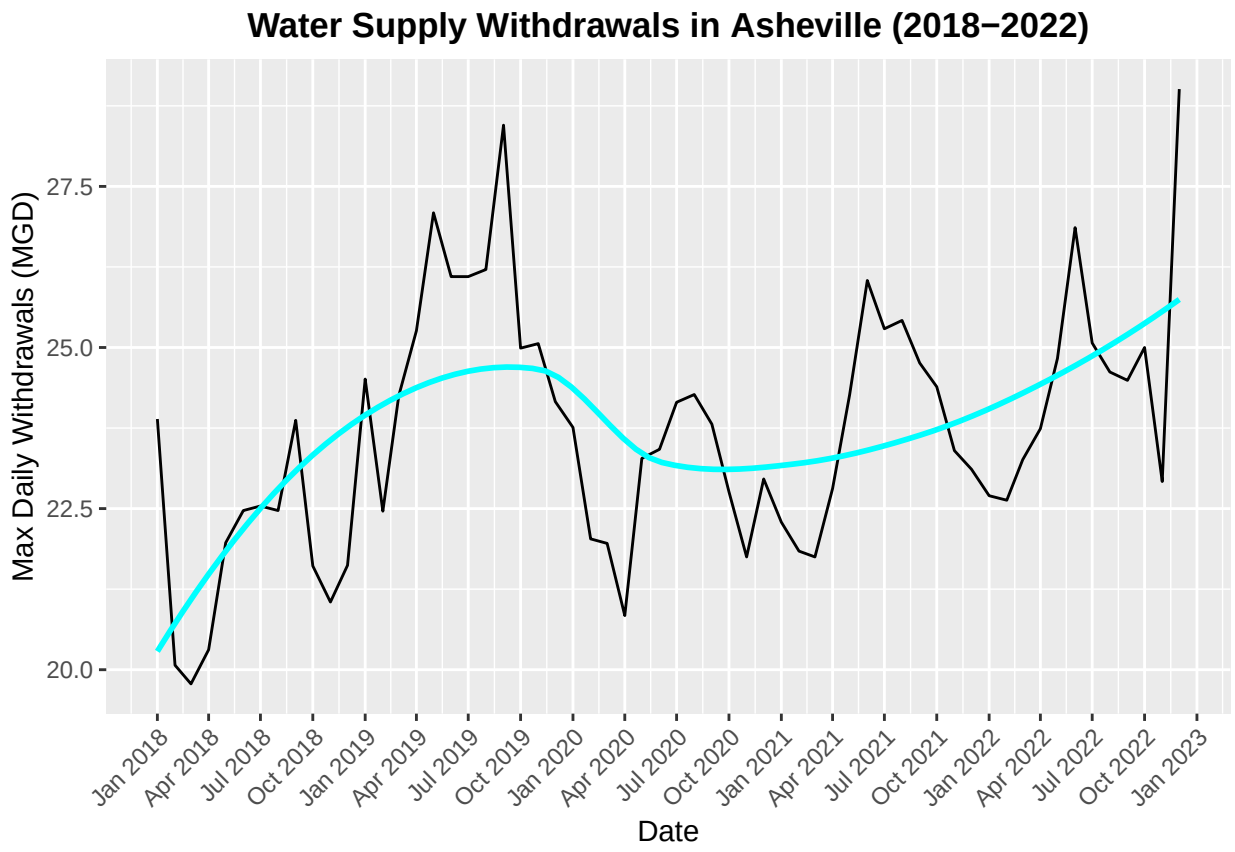
9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2018 thru 2022.Add a smoothed line to the plot (method = 'loess').

   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

years <- 2018:2022

df.list <- map2(rep('01-11-010', length(years)), years, scrape.fn)

df.asheville.2018.2022 <- bind_rows(df.list)

ggplot(df.asheville.2018.2022, aes(x = Date, y = Max.Day.Use)) +
  geom_line() +
  geom_smooth(method = "loess", se = FALSE, color = "cyan") +
  labs(
```

```
    title = "Water Supply Withdrawals in Asheville (2018-2022)",
    x = "Date",
    y = "Max Daily Withdrawals (MGD)"
  ) +
  scale_x_date(date_labels = "%b %Y", date_breaks = "3 months") +
  theme_gray() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title = element_text(hjust = 0.5, face = "bold")
  )
```

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



Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time? > Answer: >According to the plot, it appears there is a slight increase in water usage from 2018 to 2022, with a dip starting in early 2020 and lasting until around summer 2021, after which there is a very slight increase over time.