Assignment 09: Data Scraping

Li Jia Go

OVERVIEW

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

Directions

1. Rename this file `<FirstLast>_A09_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 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 packages `tidyverse`, `rvest`, and any others you end up using.
- Set your ggplot theme

```
# 1

# checking working directory
getwd()
```

```
## [1] "/home/guest/R/EDA-Fall2022"
```

```
# loading relevant packages
library(tidyverse)
library(rvest)
library(dplyr)

# setting ggplot theme
mytheme <- theme_classic(base_size = 14) + theme(axis.text = element_text(color = "black"),
    legend.position = "top", plot.title = element_text(size = 12))
theme_set(mytheme)
```

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham's 2021 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=2021

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

```
# 2
DEQ_website <- read_html("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2021")
```

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

- From the "1. System Information" section:
- Water system name
- PSWID
- Ownership
- From the "3. Water Supply Sources" section:
- Maximum Daily 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), with the first value being "27.6400".

```
# 3
water.system.name <- DEQ_website %>%
    html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
    html_text()

pswid <- DEQ_website %>%
    html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
    html_text

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

max.withdrawals.mgd <- DEQ_website %>%
    html_nodes("th~ td+ td") %>%
    html_text
```

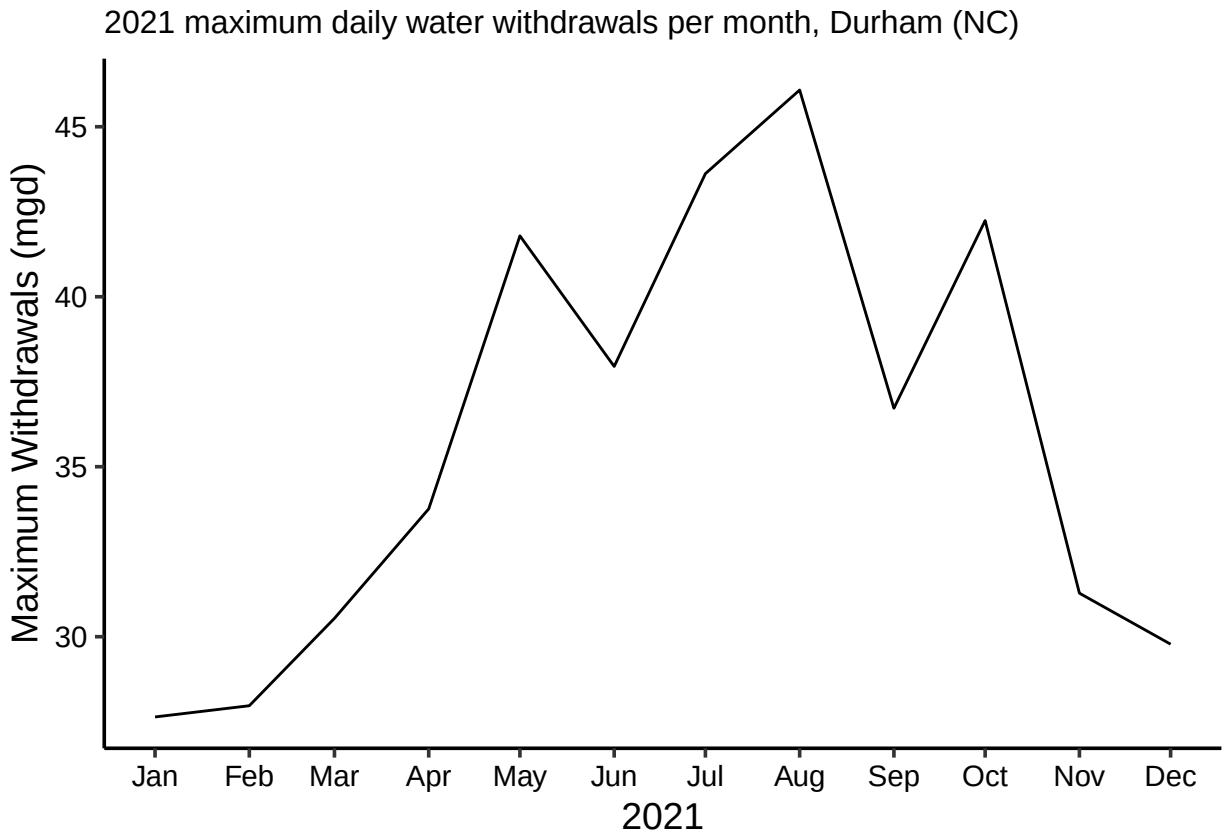
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   > 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...

5. Create a line plot of the maximum daily withdrawals across the months for 2021

```
# 4

# creating dataframe and month in order of scrape rearranging month in
# chronological order
df_withdrawals <- data.frame(Month = c(1, 5, 9, 2, 6, 10, 3, 7, 11, 4, 8, 12), Year = rep(2021,
    12), Max_Withdrawals_mgd = as.numeric(max.withdrawals.mgd)) %>%
    mutate(Date = lubridate::my(paste(Month, "-", Year))) %>%
    dplyr::arrange(Date)
```

```r
# 5 line plot of maximum daily withdrawals across months for 2021
ggplot(df_withdrawals, aes(x = Date, y = Max_Withdrawals_mgd)) + geom_line(aes()) +
    scale_x_date(date_breaks = "1 month", date_labels = "%b") + ylab("Maximum Withdrawals (mgd)") +
    xlab("2021") + ggtitle("2021 maximum daily water withdrawals per month, Durham (NC)")
```



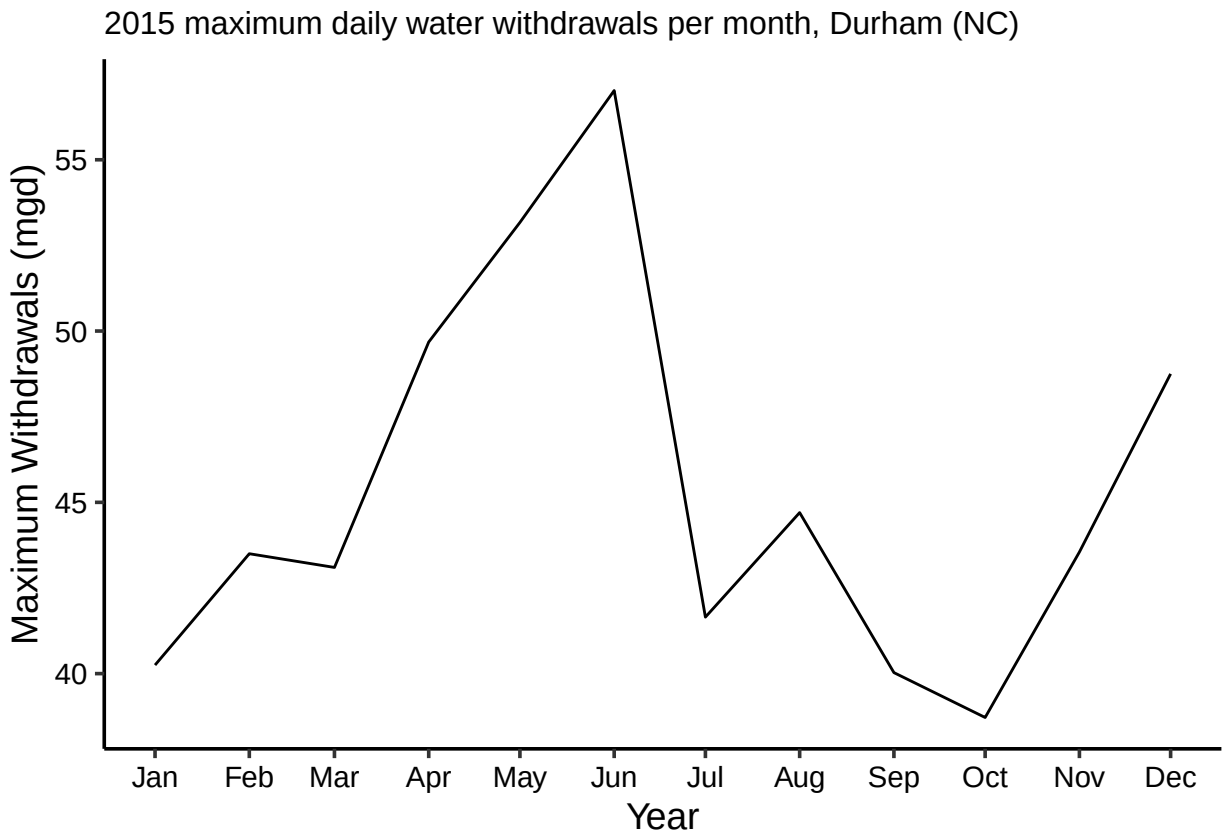
6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

```r
# 6.

# creating function
scrape.it <- function(the_year, the_PWSID) {

    # Retrieving the website contents
    theDEQsite <- read_html(paste0("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=",
        the_PWSID, "&year=", the_year))

    # locating elements and reading text attributes into variables
    water.system.name <- theDEQsite %>%
        html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
        html_text()
    pswid <- theDEQsite %>%
        html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
        html_text
```

```
    ownership <- theDEQsite %>%
        html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%
        html_text
    max.withdrawals.mgd <- theDEQsite %>%
        html_nodes("th~ td+ td") %>%
        html_text

    # creating the dataframe
    df_withdrawals_func <- data.frame(Month = c(1, 5, 9, 2, 6, 10, 3, 7, 11, 4, 8,
        12), Year = rep(the_year, 12), Max_Withdrawals_mgd = as.numeric(max.withdrawals.mgd)) %>%
        mutate(WaterSystemName = !!water.system.name, PSWID = !!pswid, Ownership = !!ownership,
            Date = lubridate::my(paste(Month, "-", Year))) %>%
        dplyr::arrange(Date)

    # Return the dataframe
    return(df_withdrawals_func)
}
```

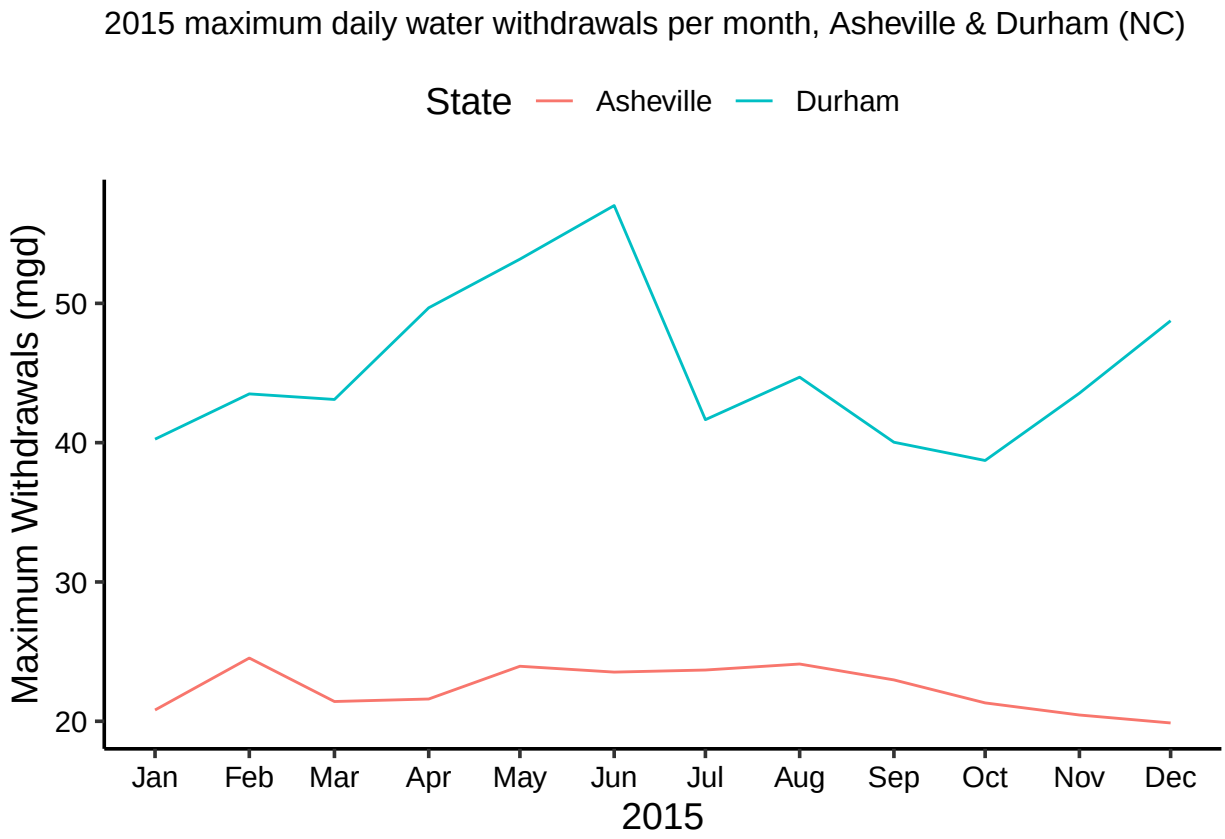
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID=‘03-32-010’) for each month in 2015

```
# 7 scraping for Durham, 2015
Durham_2015 <- scrape.it(2015, "03-32-010")

# plotting max daily withdrawals for Durham, 2015
ggplot(Durham_2015, aes(x = Date, y = Max_Withdrawals_mgd)) + geom_line(aes()) +
    scale_x_date(date_breaks = "1 month", date_labels = "%b") + ylab("Maximum Withdrawals (mgd)") +
    xlab("Year") + ggtitle("2015 maximum daily water withdrawals per month, Durham (NC)")
```



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.

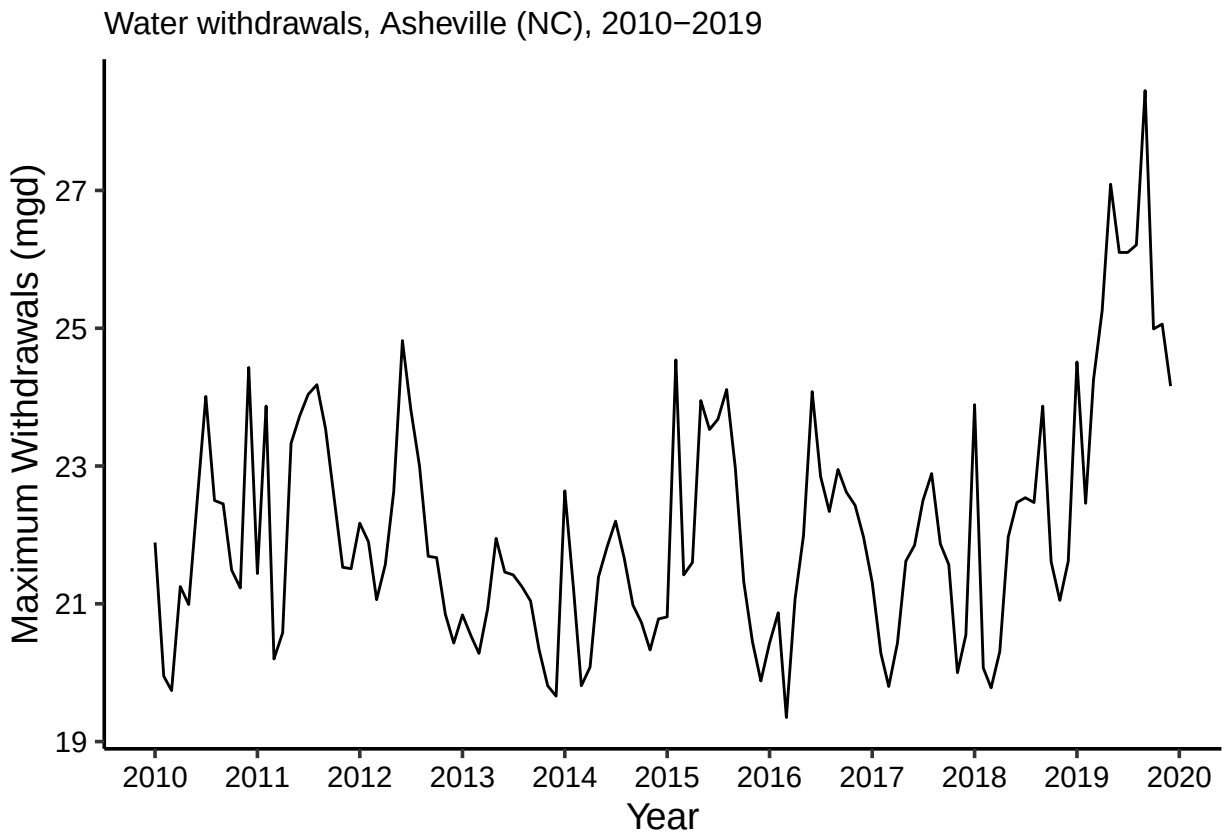
```
# 8
Asheville_2015 <- scrape.it(2015, "01-11-010")

# merging both datasets into one
# dataframe
AshevilleDurham_2015 <- rbind(Asheville_2015,
    Durham_2015)

# plotting Asheville and Durham's water
# withdrawals
ggplot(AshevilleDurham_2015, aes(x = Date,
    y = Max_Withdrawals_mgd)) + geom_line(aes(color = WaterSystemName)) +
    scale_x_date(date_breaks = "1 month",
        date_labels = "%b") + labs(x = "2015",
    y = "Maximum Withdrawals (mgd)", color = "State") +
    ggtitle("2015 maximum daily water withdrawals per month, Asheville & Durham (NC)")
```



9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

   TIP: See Section 3.2 in the "09_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

decade <- rep(2010:2019)
AshevillePWSID <- "01-11-010"

df.Asheville.decade <- map(decade, scrape.it, the_PWSID = AshevillePWSID)

df.Asheville.decade <- bind_rows(df.Asheville.decade)

# plotting Asheville's water withdrawals, 2010-2019
ggplot(df.Asheville.decade, aes(x = Date, y = Max_Withdrawals_mgd)) + geom_line(aes()) +
    scale_x_date(date_breaks = "1 year", date_labels = "%Y") + ylab("Maximum Withdrawals (mgd)") +
    xlab("Year") + ggtitle("Water withdrawals, Asheville (NC), 2010-2019")
```



Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time? Answer: The usage of water in Asheville was relatively stable from 2010-2018, with lower usage from 2013-2014. From 2019 onwards, there was a significant increase in usage of water.