Assignment 09: Data Scraping

Logan Dye

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
getwd()
```

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

```
library(tidyverse)
library(rvest)
library(lubridate)

mytheme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "red"),
        legend.position = "bottom")
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

the_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 <- the_website %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()

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

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

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

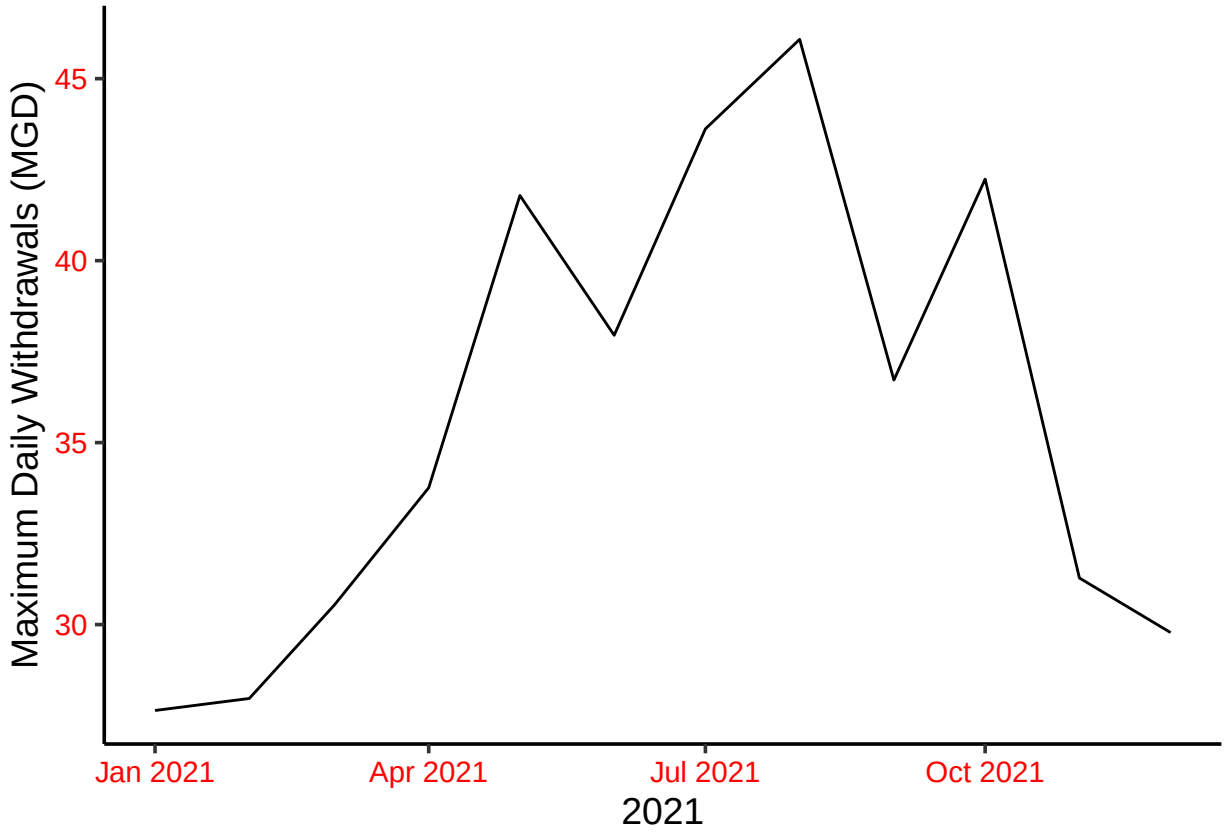
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   > 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
LWSP_df <- data.frame(
    "water.system.name" = water.system.name,
  "pwsid" = pwsid,
  "ownership" = ownership,
  "max.withdrawals.mgd" = as.numeric(max.withdrawals.mgd)
) %>%
```

```
   mutate(Month = c("Jan", "May", "Sep", "Feb", "June", "Oct", "Mar", "Jul", "Nov", "Apr", "Aug", "Dec"
          Year = rep(2021),
          Date = my(paste(Month,"-",Year)))



#5

ggplot(LWSP_df, aes(x = Date, y = max.withdrawals.mgd)) +
  geom_line() +
  xlab("2021") +
  ylab("Maximum Daily Withdrawals (MGD)")
```



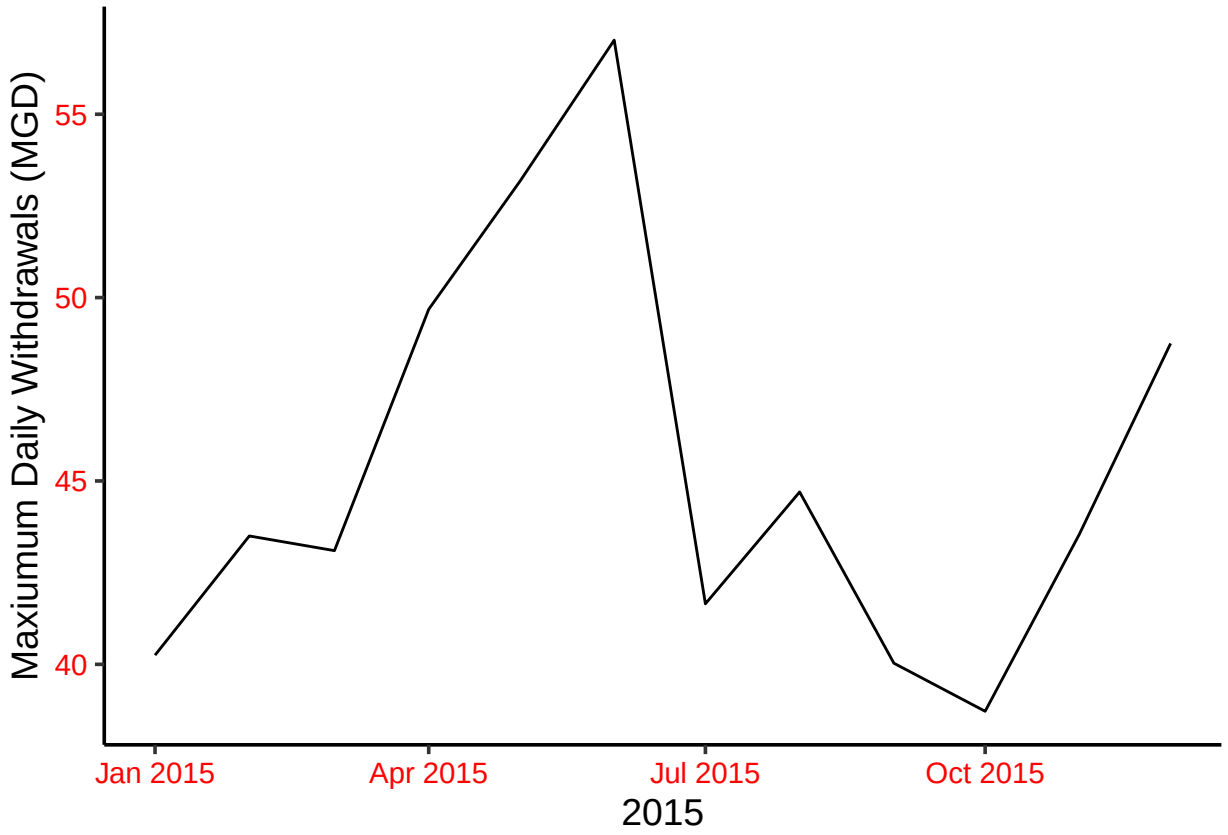
6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

```
#6.


scrape.it <- function(the_year, pwsid){

   the_website <- read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',
                                   pwsid, '&year=', the_year))

water.system.name <- the_website %>%
```

```
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()

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

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

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

  Scraped.LWSP_df <- data.frame(
    "water.system.name" = water.system.name,
  "pwsid" = pwsid,
  "ownership" = ownership,
  "max.withdrawals.mgd" = as.numeric(max.withdrawals.mgd)
) %>%
   mutate(Month = c("Jan", "May", "Sep", "Feb", "June", "Oct", "Mar", "Jul", "Nov", "Apr", "Aug", "Dec")
          Year = rep(the_year),
          Date = my(paste(Month,"-",Year)))

  return(Scraped.LWSP_df)
}
```

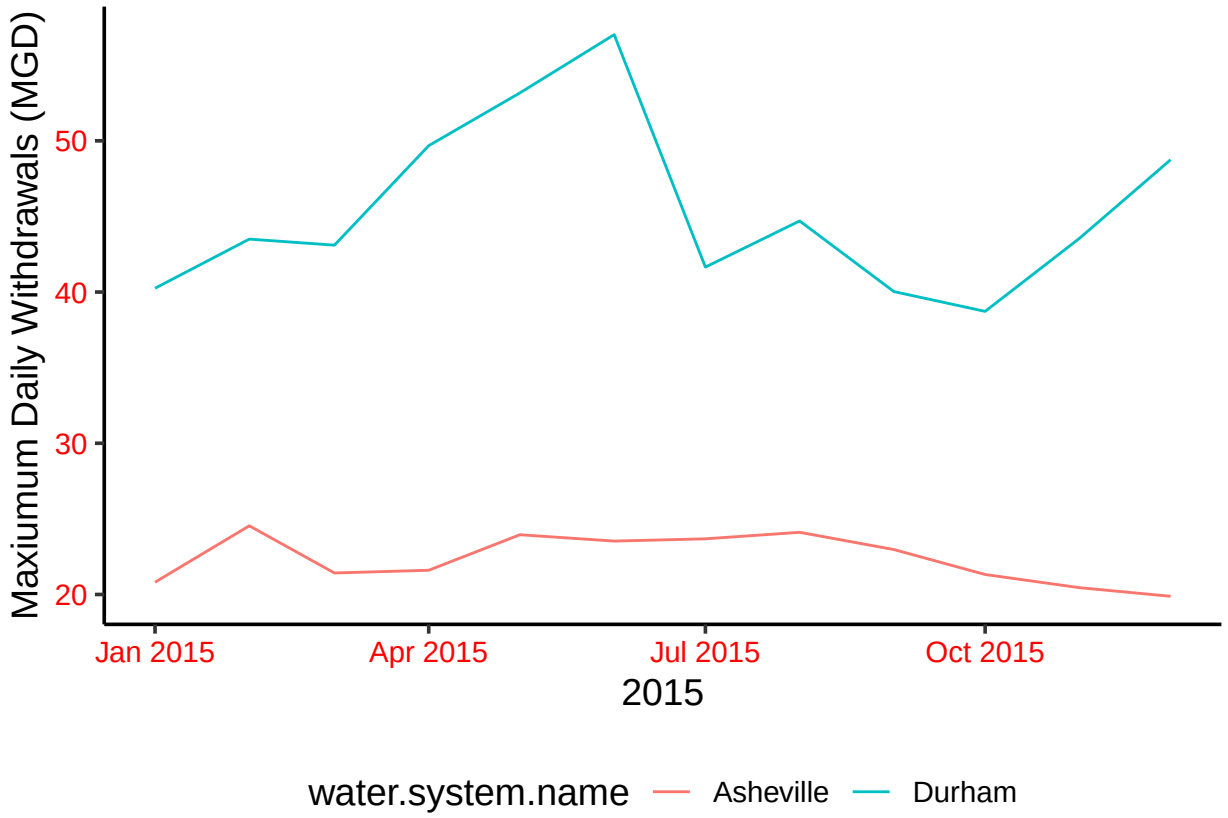
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID=‘03-32-010’) for each month in 2015

```
#7

Durham.2015.df <- scrape.it(2015, '03-32-010')

ggplot(Durham.2015.df, aes(x = Date, y = max.withdrawals.mgd)) +
  geom_line() +
  xlab("2015") +
  ylab("Maxiumum Daily Withdrawals (MGD)")
```



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.df <- scrape.it(2015, '01-11-010')

DUR.AVL.2015.df <- rbind(Asheville.2015.df, Durham.2015.df)

ggplot(DUR.AVL.2015.df, aes(x = Date, y = max.withdrawals.mgd, color = water.system.name)) +
  geom_line() +
  xlab("2015") +
  ylab("Maxiumum Daily Withdrawals (MGD)")
```

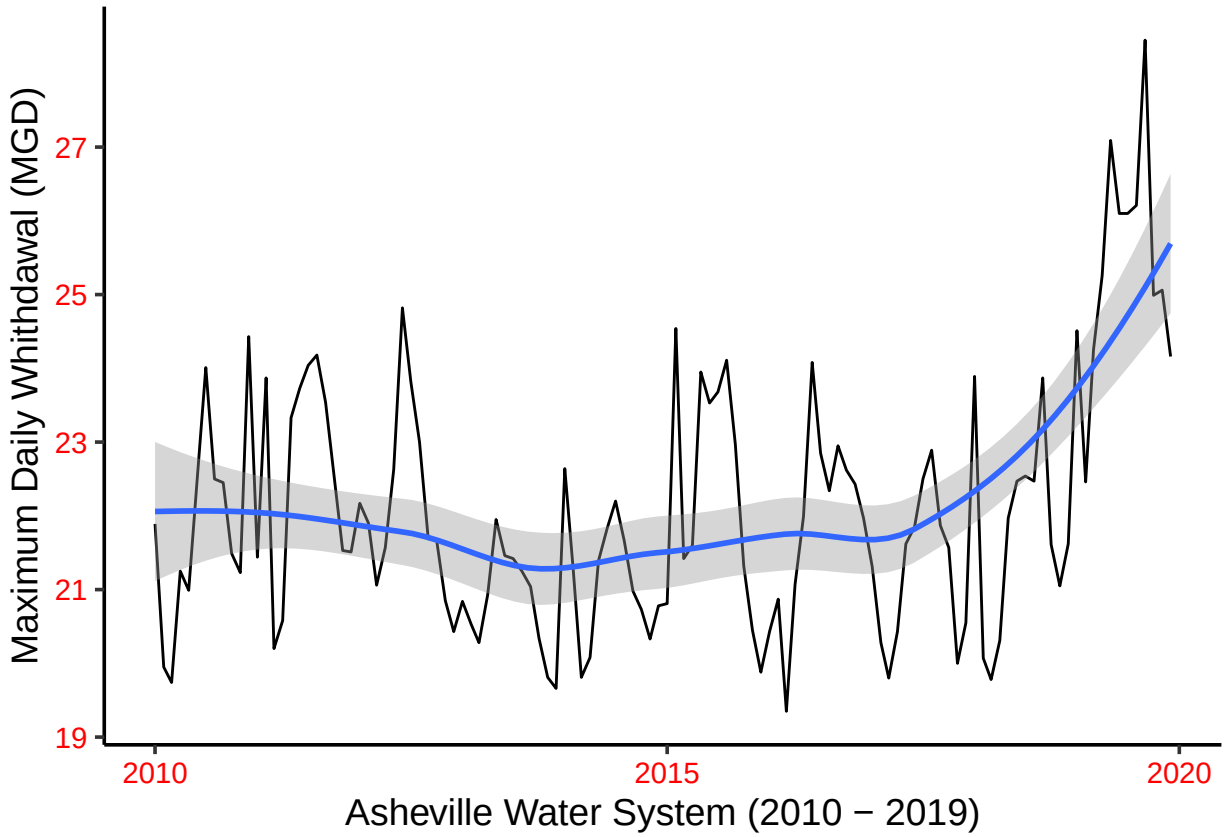


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

   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

the_years <- c(2010:2019)

AVL_10.19_df <- map2(the_years, '01-11-010', scrape.it) %>%
  bind_rows()

ggplot(AVL_10.19_df, aes(x = Date, y = max.withdrawals.mgd)) +
  geom_line() +
  geom_smooth() +
  xlab("Asheville Water System (2010 - 2019)") +
  ylab("Maximum Daily Whithdawal (MGD)")
```

```
## `geom_smooth()` using method = 'loess' and 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?

It appears that, starting around 2017, Asheville has an upward trend in its daily maximum withdawals.