Assignment 10: Data Scraping

Sammy DiLoreto

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 Set up session
library(tidyverse)
library(lubridate)
library(here)
library(rvest)
getwd()
```

```
## [1] "/Users/sammydiloreto/Library/CloudStorage/Box-Box/ENV872-EDA/EDA-Spring2023/Assignments"
```

```
here()
```

```
## [1] "/Users/sammydiloreto/Library/CloudStorage/Box-Box/ENV872-EDA/EDA-Spring2023"
```

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham's 2022 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=2022

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

```
#2 set variables for components of url
the_base_url <- 'https://www.ncwater.org/WUDC/app/LWSP/report.php'
the_pwsid <- '03-32-010'
the_year <- 2022
the_scrape_url <- paste0(the_base_url, '?pwsid=', the_pwsid, '&year=', the_year)
print(the_scrape_url)
```

```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2022"
```

```
the_website <- read_html(the_scrape_url)
```

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), with the first value being "27.6400".

```
#3 assign variables to the values we want to scrape
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()
```

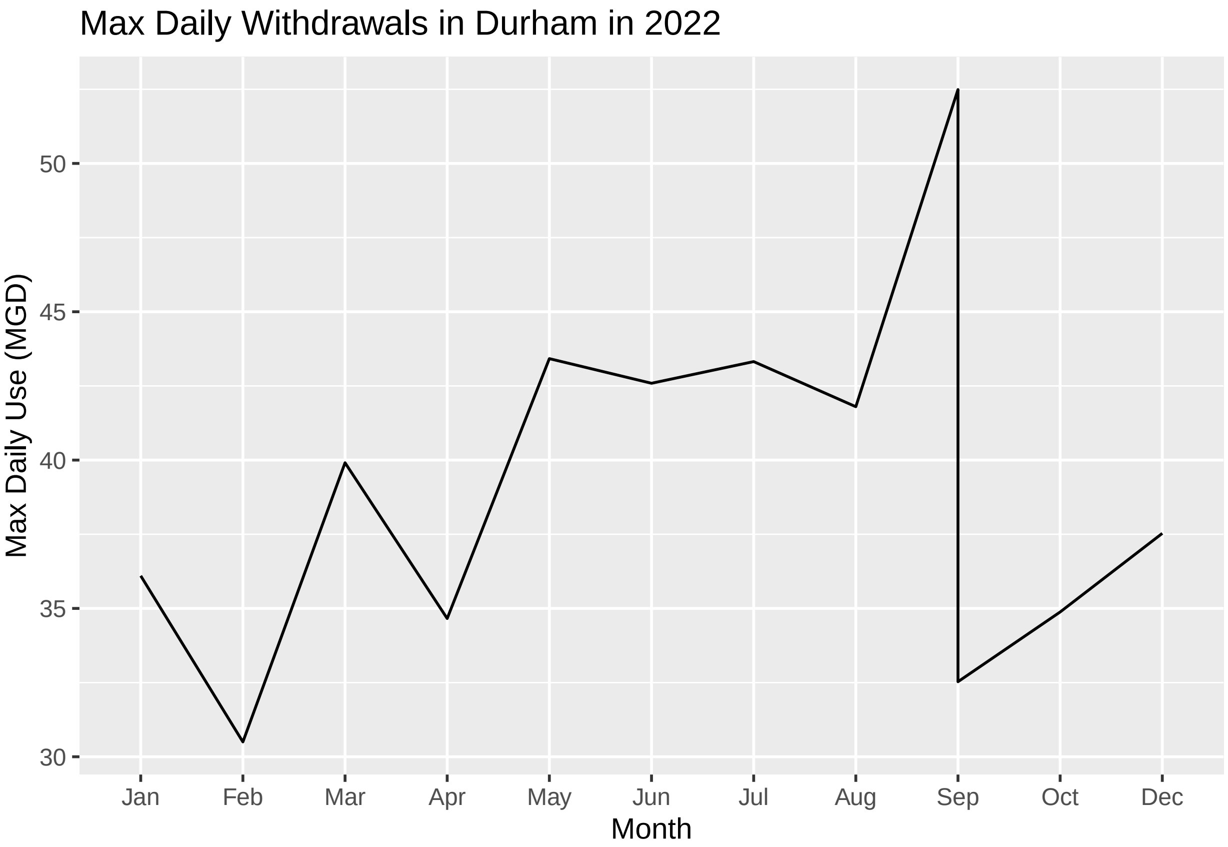
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   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… Or, you could scrape month values from the web page…

5. Create a line plot of the average daily withdrawals across the months for 2022

```r
#4 convert scraped data into dataframe
durham.2022.df <- data.frame("Month" = c(1, 5, 9, 2, 6, 10, 3, 7, 9, 4, 8, 12),
                              "Year" = the_year,
                             "Max_Withdrawals_mgd" = as.numeric(max.withdrawals.mgd)) %>%
  mutate(WaterSystemName = !!water.system.name,
         PWSID = !!PWSID,
         Ownership = !!ownership,
         Date = my(paste(Month,"-",Year)))

#5 Create a line plot of the max daily withdrawals across the months for 2022
Max.Daily.2022 <- ggplot(durham.2022.df,
       aes(x= factor(Month, levels = 1:12, labels = month.abb),
         y=Max_Withdrawals_mgd,
         group=1)) +
  geom_line() +
  labs(x = "Month",
       y="Max Daily Use (MGD)",
       title = "Max Daily Withdrawals in Durham in 2022" )
Max.Daily.2022
```

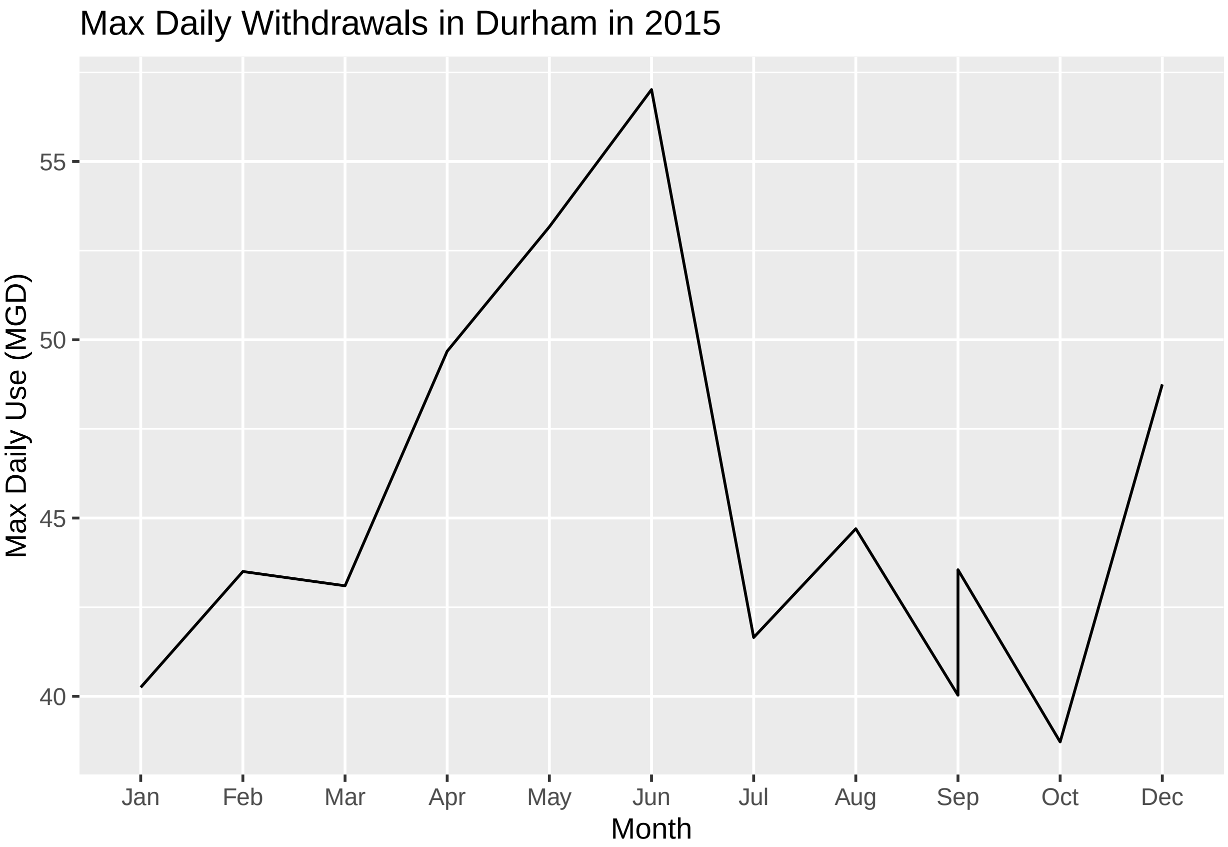


6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

```
#6. Construct function
#Create our scraping function
scrape.it <- function(the_year, the_pwsid){

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

  #Scrape the data items
  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()

  #Convert to a dataframe
  water.supply.df <- data.frame("Month" = c(1, 5, 9, 2, 6, 10, 3, 7, 9, 4, 8, 12),
                              "Year" = the_year,
                             "Max_Withdrawals_mgd" = as.numeric(max.withdrawals.mgd)) %>%
  mutate(WaterSystemName = !!water.system.name,
         PWSID = !!PWSID,
         Ownership = !!ownership,
         Date = my(paste(Month,"-",Year)))

  #Return the dataframe
  return(water.supply.df)
}
```

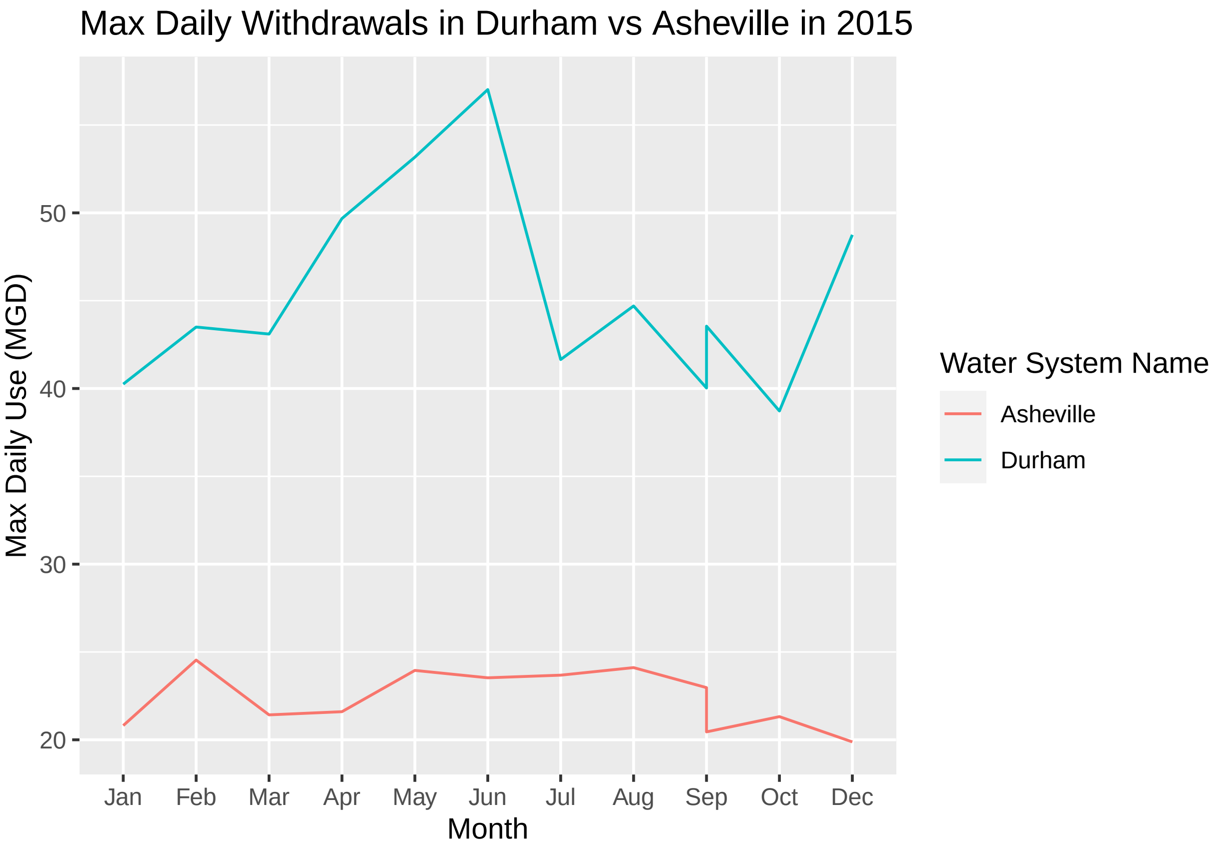
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID='03-32-010') for each month in 2015

```
#7
#extract max daily withdrawals for Durham for each month in 2015
durham.2015.df <- scrape.it(2015,'03-32-010')

#plot it
Max.Daily.2015 <- ggplot(durham.2015.df,
       aes(x= factor(Month, levels = 1:12, labels = month.abb),
         y=Max_Withdrawals_mgd,
         group = 1)) +
  geom_line() +
  labs(x = "Month",
       y="Max Daily Use (MGD)",
       title = "Max Daily Withdrawals in Durham in 2015")
Max.Daily.2015
```



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.

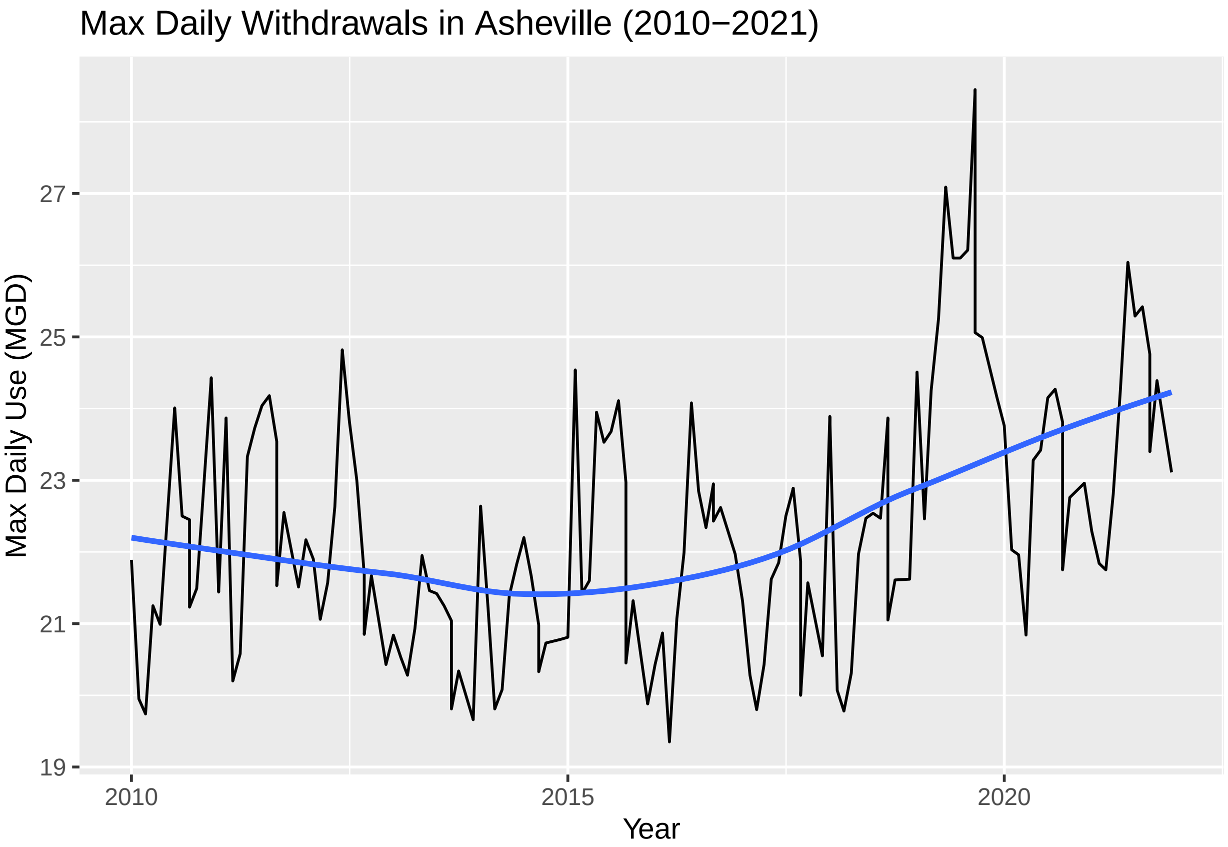
```
#8
#extract data for Asheville (PWSID = 01-11-010) in 2015
asheville.2015.df <- scrape.it(2015,'01-11-010')

#combine with Durham 2015 data
ash.durm.2015.df <- rbind(durham.2015.df, asheville.2015.df)

#plot as separate lines to compare
ggplot(ash.durm.2015.df,
       aes(x= factor(Month, levels = 1:12, labels = month.abb),
           y= Max_Withdrawals_mgd,
         color= WaterSystemName,
         group= WaterSystemName)) +
  geom_line() +
  labs(x = "Month",
       y="Max Daily Use (MGD)",
       title = "Max Daily Withdrawals in Durham vs Asheville in 2015",
       color= "Water System Name")
```



9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2021. Add a smoothed line to the plot (method = 'loess').

   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
#Set the inputs to scrape years 2015 to 2020 for the site "0004-0001"
the_years = rep(2010:2021)
my_pwsid = '01-11-010'

#Use lapply to apply the scrape function
ash.2010.2021.dfs <- lapply(X = the_years,
                  FUN = scrape.it,
                  the_pwsid=my_pwsid)

#Conflate the returned dataframes into a single dataframe
ash.2010.2021.df <- bind_rows(ash.2010.2021.dfs)

#plot it
ggplot(ash.2010.2021.df,aes(x=Date,y=Max_Withdrawals_mgd)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE)+
  labs(x = "Year",
       y="Max Daily Use (MGD)",
       title = "Max Daily Withdrawals in Asheville (2010-2021)")
```



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 Asheville has been increasing their water usage since 2015. Between 2010 and 2015 they decreased their water usage slightly but have been increasing since then.