# Assignment 09: Data Scraping

Ariel Lam

## 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] "/Users/ariellam/Desktop/EDA-Fall2022/Assignments"
```

```
library(tidyverse)
library(rvest)
library(lubridate)
library(viridis)
library(dataRetrieval)
library(tidycensus)
library(formatR)
knitr::opts_chunk$set(tidy.opts=list(width.cutoff=45), tidy=TRUE)


mytheme <- theme_classic() +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
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
durham_municipality <- read_html("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=
durham_municipality
```

```
## {html_document}
## <html xmlns="http://www.w3.org/1999/xhtml" lang="en" xml:lang="en">
## [1] <head>\n<title>DWR :: Local Water Supply Planning</title>\n<meta http-equ ...
## [2] <body id="plan">\r\n<!--<div id="division-header">\r\n<a name="top" href= ...
```

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

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

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

max.withdrawals.mgd <- durham_municipality %>%
    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.)

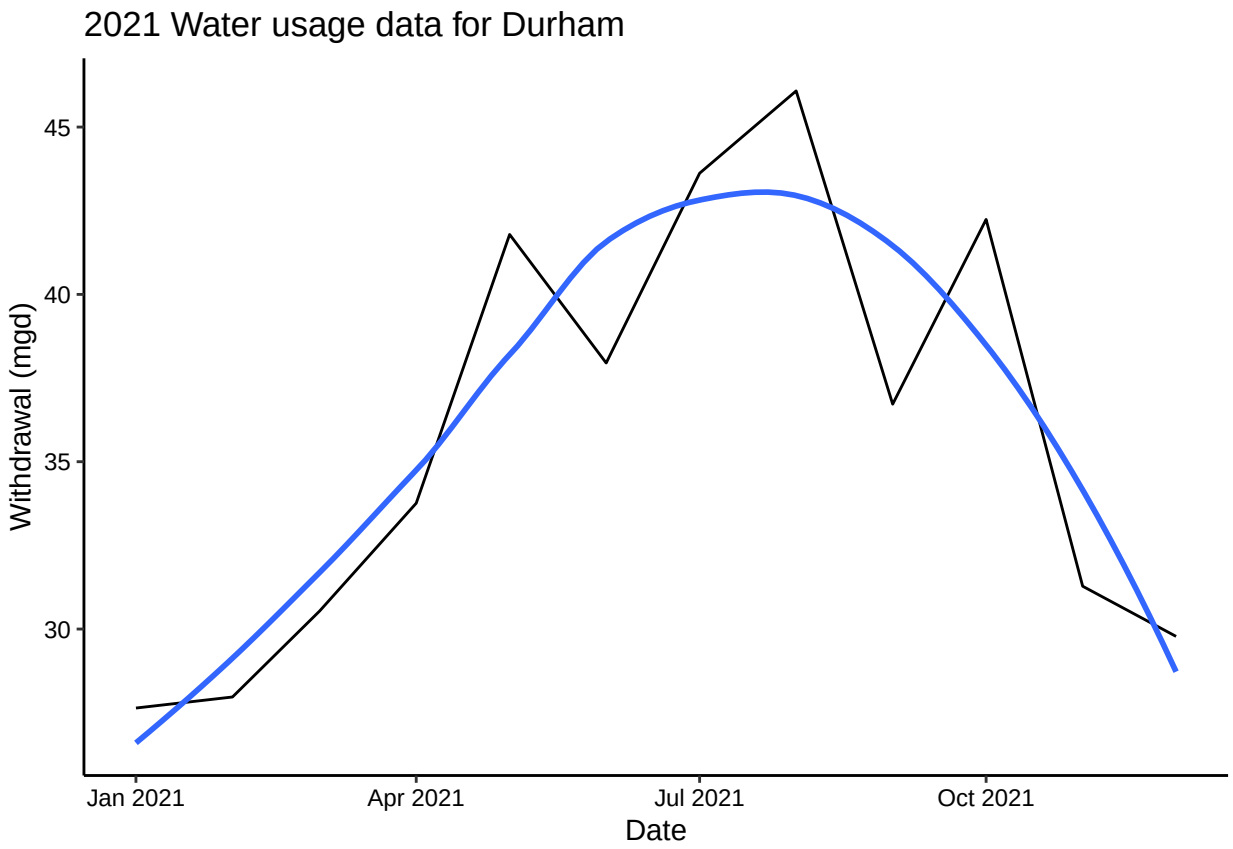
   TIP: Use `rep()` to repeat a value when creating a dataframe.

   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
durham_withdrawals <- data.frame(Name = rep(water.system.name),
    PSWID = rep(pswid), Ownership = rep(ownership),
    Month = month(c(1, 5, 9, 2, 6, 10, 3, 7, 11,
        4, 8, 12)), Year = 2021, Max_Withdrawals_mgd = as.numeric(max.withdrawals.mgd))
durham_withdrawals <- durham_withdrawals %>%
    mutate(Date = my(paste0(Month, "-", Year)))

# 5


ggplot(durham_withdrawals, aes(x = Date, y = Max_Withdrawals_mgd,
    group = 1)) + geom_line() + geom_smooth(method = "loess",
    se = FALSE) + labs(title = "2021 Water usage data for Durham",
    y = "Withdrawal (mgd)", x = "Date")
```

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




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**.

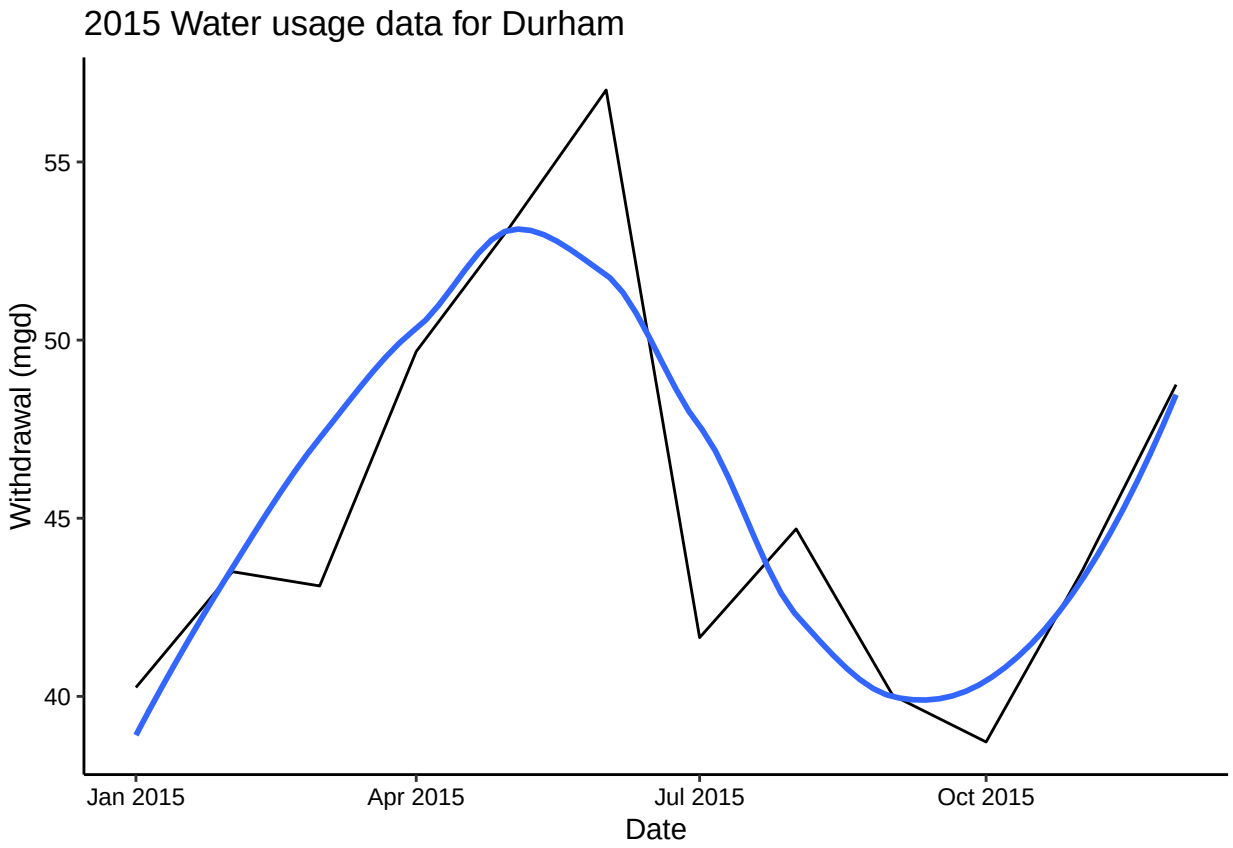
```
# 6.
scrape.it <- function(the_year, the_site) {

    # Retrieve the website contents
    the_website <- read_html(paste0("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=",
```

```
        the_site, "&year=", the_year))

    water_system_tag <- "div+ table tr:nth-child(1) td:nth-child(2)"
    site_tag <- "td tr:nth-child(1) td:nth-child(5)"
    ownership_tag <- "div+ table tr:nth-child(2) td:nth-child(4)"
    the_data_tag <- "th~ td+ td"

    # Scrape the data items
    the_water_system <- the_website %>%
        html_nodes(water_system_tag) %>%
        html_text()
    site_name <- the_website %>%
        html_nodes(site_tag) %>%
        html_text()
    ownership_type <- the_website %>%
        html_nodes(ownership_tag) %>%
        html_text()
    max_withdrawals <- the_website %>%
        html_nodes(the_data_tag) %>%
        html_text()

    # Convert to a dataframe
    df_withdrawals <- data.frame(Name = rep(the_water_system),
        PSWID = rep(site_name), Ownership = rep(ownership_type),
        Month = month(c(1, 5, 9, 2, 6, 10, 3,
            7, 11, 4, 8, 12)), Year = the_year,
        Max_Withdrawals_mgd = as.numeric(max_withdrawals))
    df_withdrawals <- df_withdrawals %>%
        mutate(Date = my(paste0(Month, "-", Year)))

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

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

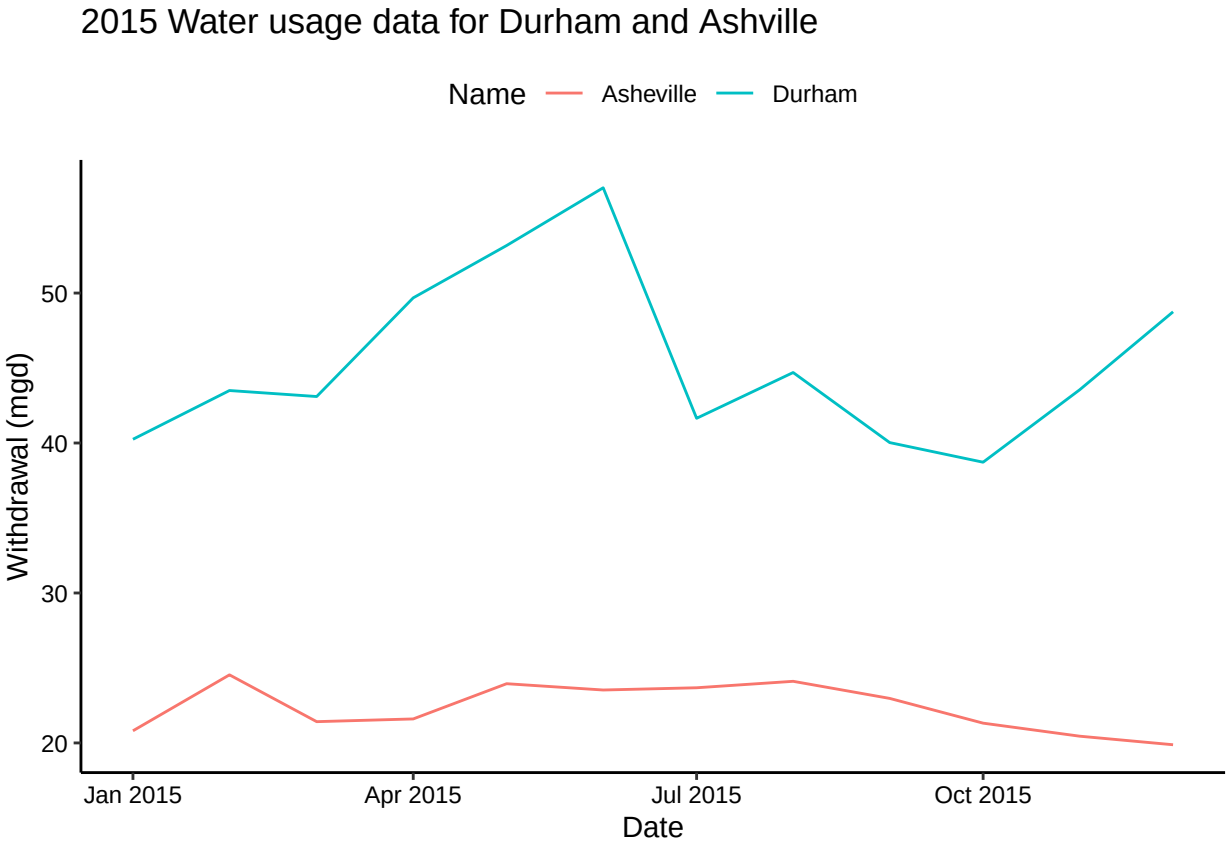
```
# 7
durham_2015 <- scrape.it(2015, "03-32-010")
view(durham_2015)

ggplot(durham_2015, aes(x = Date, y = Max_Withdrawals_mgd,
    group = 1)) + geom_line() + geom_smooth(method = "loess",
    se = FALSE) + labs(title = "2015 Water usage data for Durham",
    y = "Withdrawal (mgd)", x = "Date")
```

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




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
ashville_2015 <- scrape.it(2015, "01-11-010")
view(ashville_2015)

combine_2015 <- rbind(ashville_2015, durham_2015)


ggplot(combine_2015, aes(x = Date, y = Max_Withdrawals_mgd,
    group = Name)) + geom_line(aes(color = Name)) +
    labs(title = "2015 Water usage data for Durham and Ashville",
        y = "Withdrawal (mgd)", x = "Date")
```

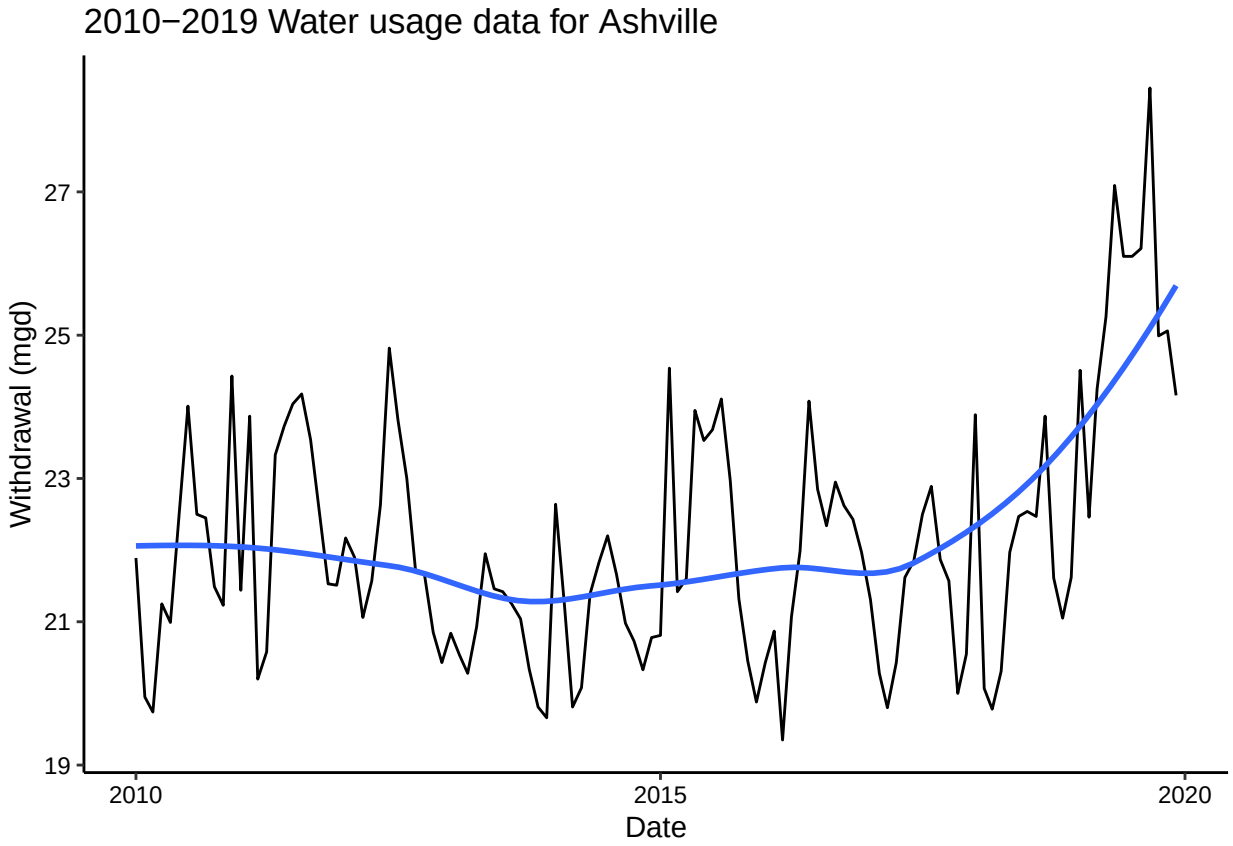



9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

   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, 2011, 2012, 2013, 2014, 2015,
    2016, 2017, 2018, 2019)

asheville_2010_2019 <- map2(the_years, "01-11-010",
    scrape.it) %>%
    bind_rows()

# asheville_2010_2019_dataframe <-
# bind_rows(asheville_2010_2019)

ggplot(asheville_2010_2019, aes(x = Date, y = Max_Withdrawals_mgd)) +
    geom_line() + geom_smooth(method = "loess",
    se = FALSE) + labs(title = "2010-2019 Water usage data for Ashville",
    y = "Withdrawal (mgd)", x = "Date")
```

```
## `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? The water usage for Ashville seems to be gradually increasing over time.