# Assignment 09: Data Scraping

Yanxi Peng

## 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

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

getwd()
```

```
## [1] "C:/Users/16920/Documents/R/EDA-Fall2022"
```

```
setwd('C:/Users/16920/Documents/R/EDA-Fall2022')

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

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

```
## {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".

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

```
## [1] "Durham"
```

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

```
## [1] "03-32-010"
```

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

```
## [1] "Municipality"
```

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

```
##  [1] "27.6400" "41.7900" "36.7200" "27.9700" "37.9500" "42.2400" "30.5400"
##  [8] "43.6200" "31.2800" "33.7600" "46.0800" "29.7800"
```

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.

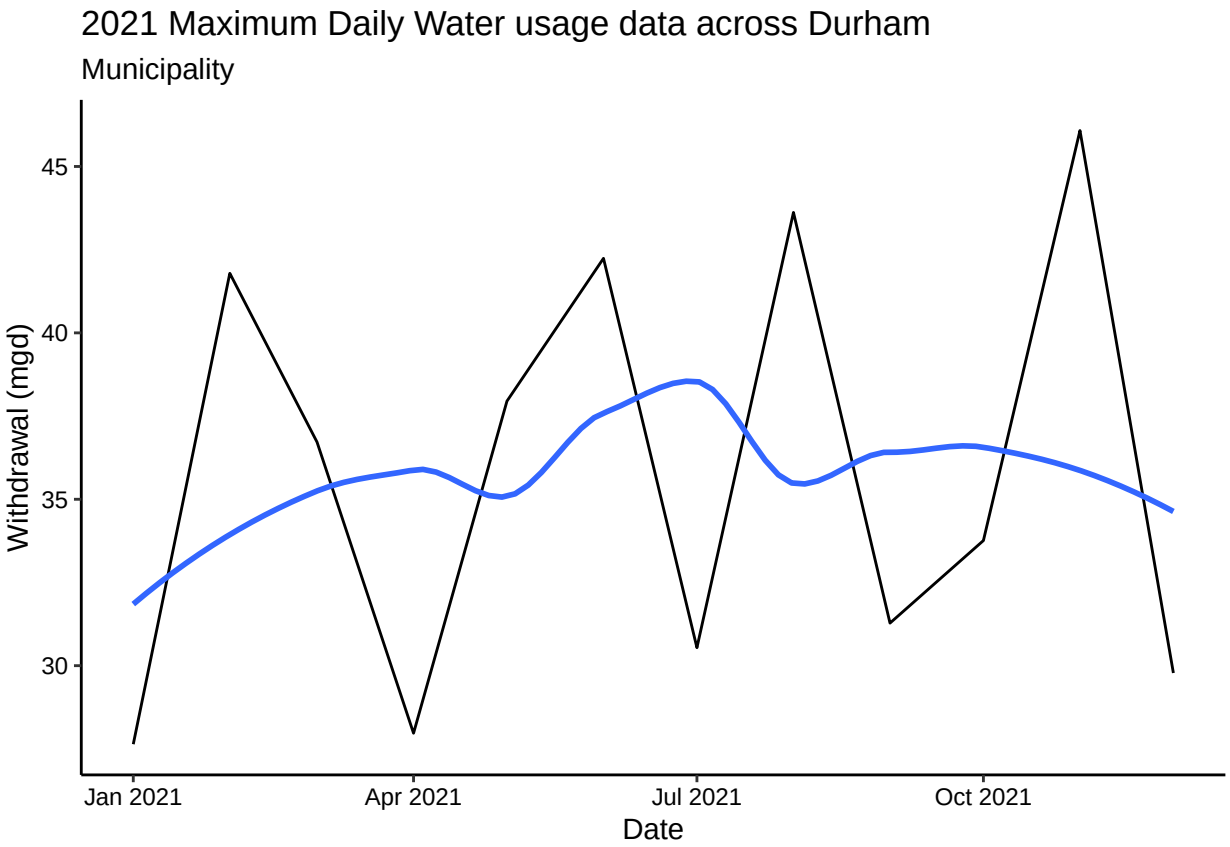
   NOTE: It's likely you won't be able to scrape the monthly widthrawal data in chronological order. You can overcome this by creating a month column manually assigning values in the order the data are scraped: "Jan", "May", "Sept", "Feb", etc...

```
df_withdrawals <- data.frame("Month" = rep(1:12),
                             "Year" = rep(2021,12),
                             "Max_Withdrawals_mgd" = as.numeric(max.withdrawals.mgd))
df_withdrawals <- df_withdrawals %>%
  mutate(Water_System_Name = !!water.system.name,
         PSWID = !!pswid,
         Ownership = !!ownership,
         Date = my(paste(Month,"-",Year)))
```

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

```
ggplot(df_withdrawals,aes(x=Date,y=Max_Withdrawals_mgd)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2021 Maximum Daily Water usage data across",water.system.name),
       subtitle = ownership,
       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**.

```
scrape.it <- 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))

  the_water_system_name_tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'
  the_pswid_tag <- 'td tr:nth-child(1) td:nth-child(5)'
  the_ownership_tag <- 'div+ table tr:nth-child(2) td:nth-child(4)'
  the_max_withdrawals_mgd_tag <- 'th~ td+ td'

  the_water_system_name <- the_website %>%
    html_nodes(the_water_system_name_tag) %>% html_text()
  the_pswid_name <- the_website %>%
    html_nodes(the_pswid_tag) %>%  html_text()
  the_ownership_name <- the_website %>%
    html_nodes(the_ownership_tag) %>% html_text()
  max_withdrawals <- the_website %>%
    html_nodes(the_max_withdrawals_mgd_tag) %>% html_text()

df_withdrawals <- data.frame("Month" = rep(1:12),
                             "Year" = rep(the_year,12),
```

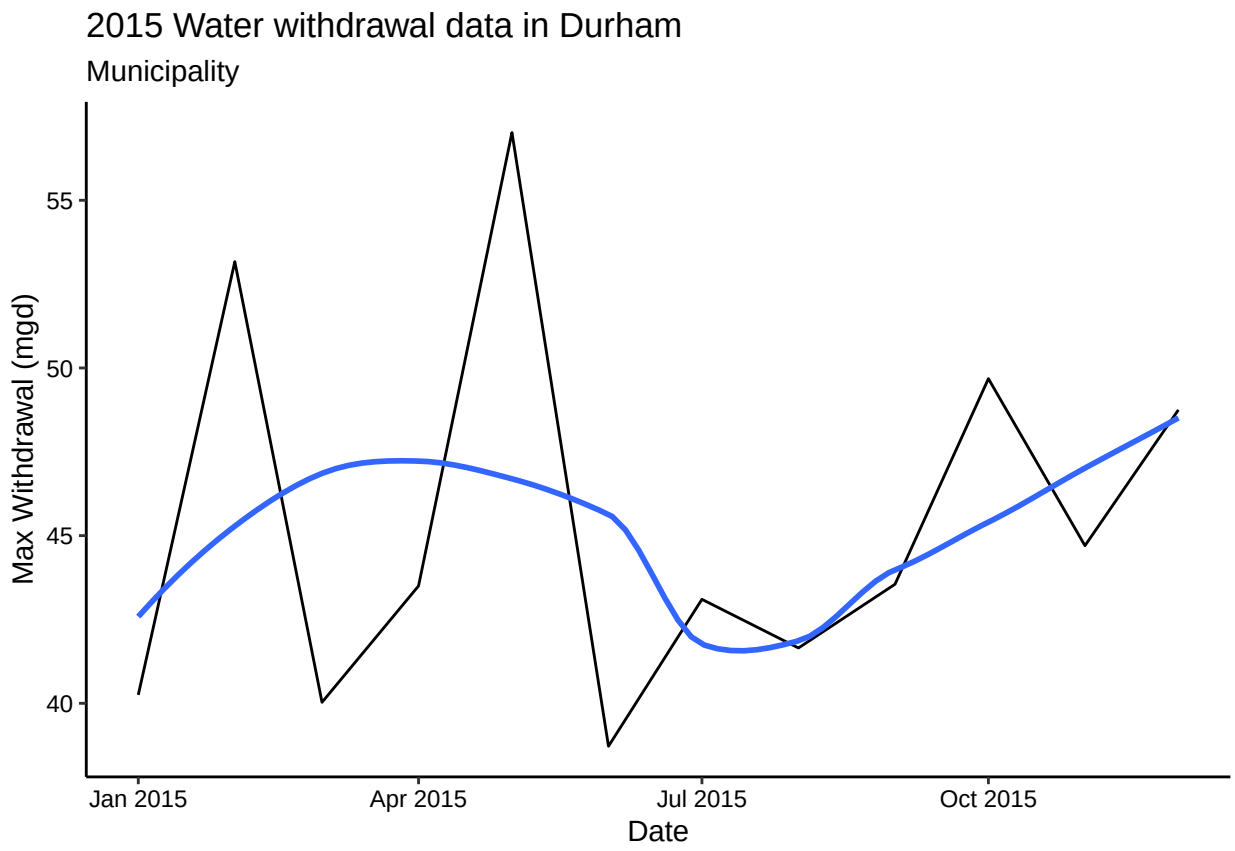
```
                              "max_withdrawals" =
                                as.numeric(max_withdrawals)) %>%
  mutate(water_system = the_water_system_name,
         pswid = !!the_pswid_name,
         ownership = !!the_ownership_name,
         Date = my(paste(Month,"-",Year)))
df_withdrawals <- arrange(df_withdrawals, Month)

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

```
Durham_2015 <- scrape.it('03-32-010','2015')
ggplot(Durham_2015,aes(x = Date, y = max_withdrawals)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2015 Water withdrawal data in Durham"),
       subtitle = 'Municipality',
       y="Max Withdrawal (mgd)",
       x="Date")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```

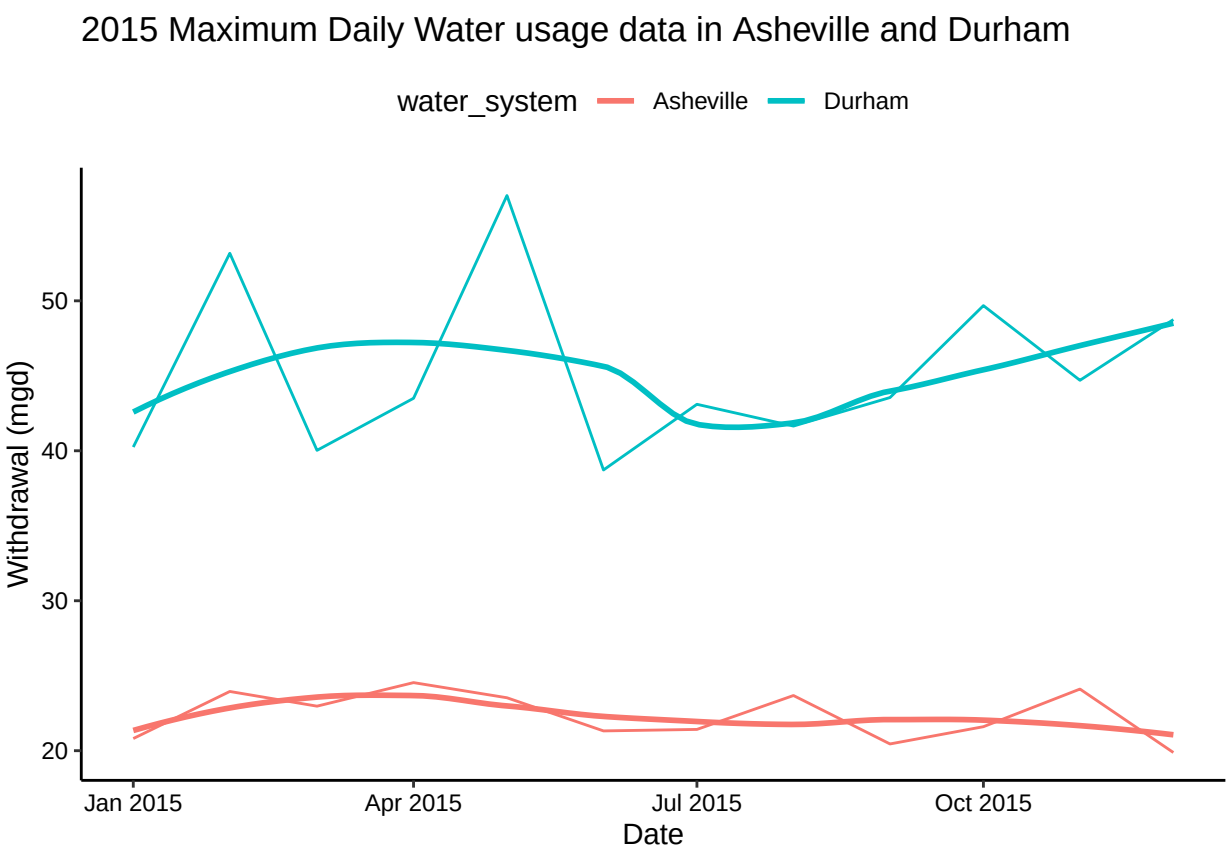


8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares Asheville's to Durham's water withdrawals.

```
Asheville_2015 <-scrape.it('01-11-010','2015')

DA2015<- rbind(Durham_2015,Asheville_2015)
ggplot(DA2015,aes(x=Date,y=max_withdrawals,
                  color=water_system,))+
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2015 Maximum Daily Water usage data in Asheville and Durham"),
       y="Withdrawal (mgd)",
       x="Date")
```

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




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.

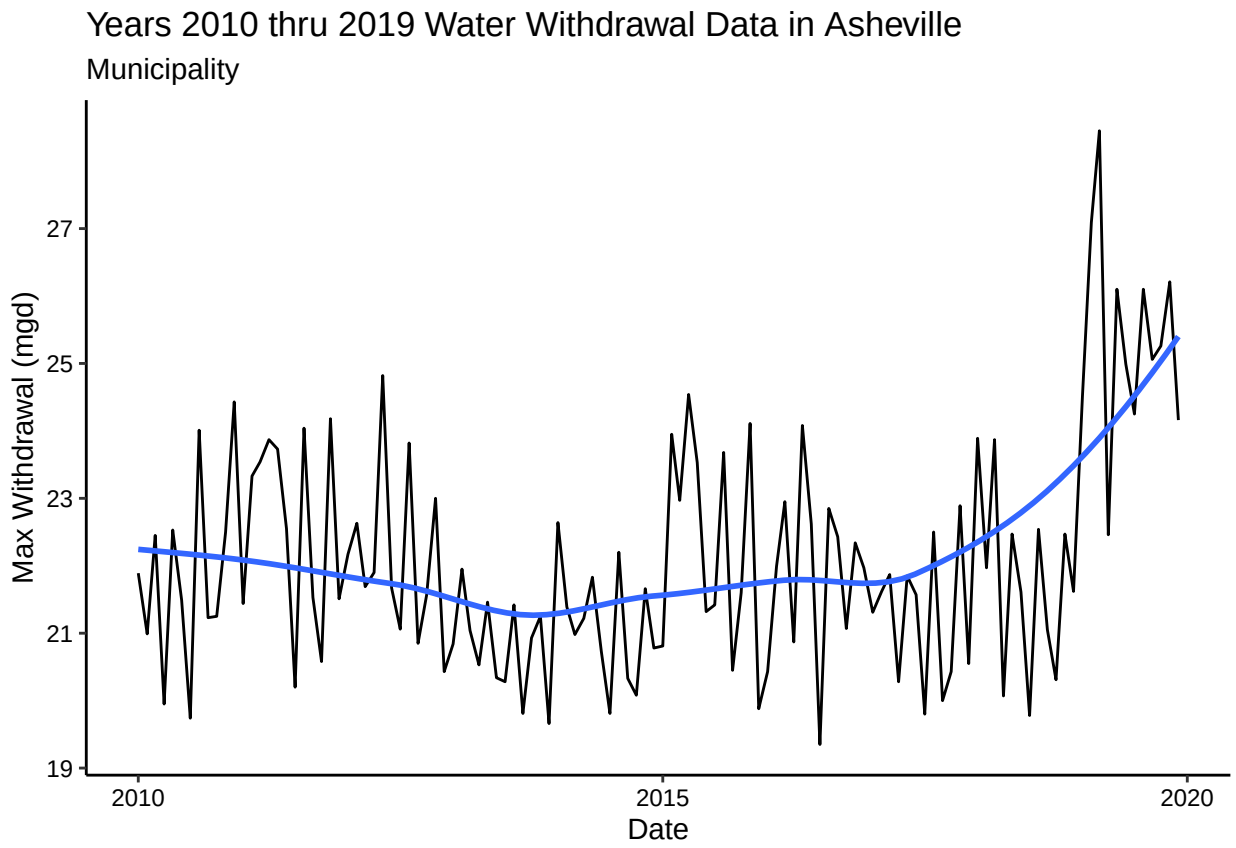
   TIP: See Section 3.2 in the "09_Data_Scraping.Rmd" where we apply "map2()" to iteratively run a function over two inputs. Pipe the output of the map2() function to `bindrows()` to combine the dataframes into a single one.

```
ash_pswid <- rep.int('01-11-010', 10)
ash_multiyears <- rep(2010:2019, 1)
ash <- map2(ash_pswid, ash_multiyears, scrape.it)

ash_date <- bind_rows(ash)

ggplot(ash_date, aes(y = max_withdrawals, x=Date)) +
  geom_line() +
  geom_smooth(method="loess", se=FALSE) +
  labs(title = paste("Years 2010 thru 2019 Water Withdrawal Data in Asheville"),
       subtitle = 'Municipality',
       y="Max 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? From the smooth line, there have a increasing trend in water usage over time in Asheville.