Assignment 10: Data Scraping

Simon Heinberg

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
library(tidyverse)
library(lubridate)
library(here); here()
```

```
## [1] "/Users/simonheinberg/Desktop/Duke_Coursework/Envirnmental_Data_Exploration/EDE_Fall2023"
```

```
library(rvest)
library(ggplot2)
```

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

lwsp_webpage <- read_html('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2022')
```

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)".

```
#3

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

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

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

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

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

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

```
MGD <- lwsp_webpage %>%
  html_nodes("th~ td+ td") %>%
  html_text()
MGD
```

```
##  [1] "36.1000" "43.4200" "52.4900" "30.5000" "42.5900" "34.8800" "39.9100"
##  [8] "43.3200" "32.5300" "34.6600" "41.8000" "37.5300"
```

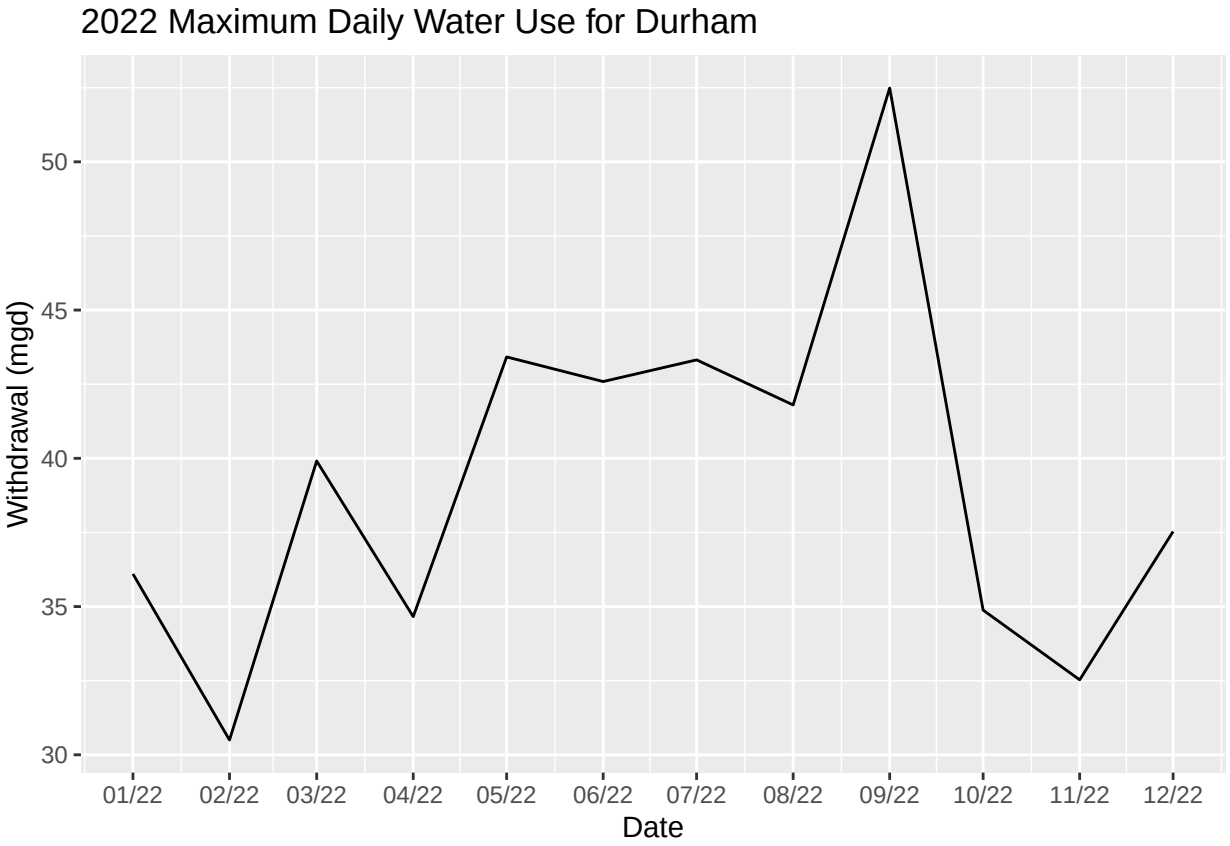
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   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 maximum daily withdrawals across the months for 2022

```
#4

MM <- lwsp_webpage %>%
  html_nodes(".fancy-table:nth-child(31) tr+ tr th") %>%
  html_text()

mgd_dataframe <- data.frame("Month" = as.factor(MM),
                            "Year" = rep(2022,12),
                            "Max Daily Use" = as.numeric(MGD)

                            )
mgd_dataframe <- mgd_dataframe %>%
  mutate(owner=!!ownership,
         ID=!!PWSID,
         Municipality=!!water_system,
         Date = my(paste(Month,"-",Year))) %>%
         arrange(Date)


#5
ggplot(mgd_dataframe, aes(x=Date, y=Max.Daily.Use))+
  geom_line()+
  labs(title = paste("2022 Maximum Daily Water Use for", water_system),
       y="Withdrawal (mgd)",
       x="Date")+
  scale_x_date(date_labels="%m/%y", date_breaks="1 month")
```



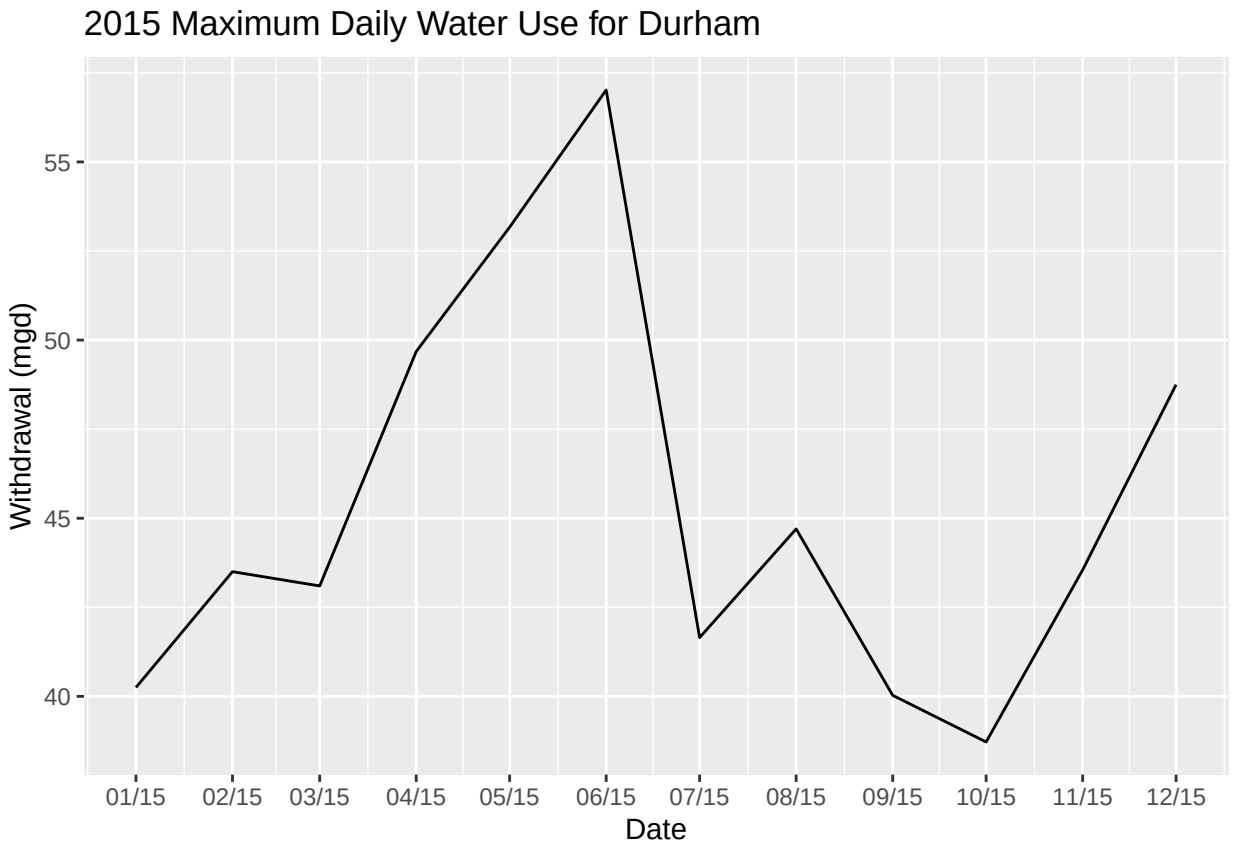
6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

```
#6.

the_year <- 2022
the_PWSID <- '03-32-010'

water_function <- function(the_year, the_PWSID){

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

   #Set the element address variables


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

 #Scrape the data items

 MM_function <- c("Jan", "May", "Sep", "Feb", "Jun", "Oct", "Mar", "Jul", "Nov", "Apr", "Aug", "Dec")
```

```
 water_system_function <- function_website %>%
  html_nodes(water_system_tag) %>%
  html_text()

PWSID_function <- function_website %>%
  html_nodes(PWSID_tag) %>%
  html_text()

ownership_function <- function_website %>%
  html_nodes(ownership_tag) %>%
  html_text()

MGD_function <- function_website %>%
  html_nodes(MGD_tag) %>%
  html_text()

 #Convert to a dataframe

function_dataframe <- data.frame("Month" = as.factor(MM_function),
                                 "Year" = rep(the_year,12),
                                 "Max Daily Use" = as.numeric(MGD_function)) %>%

  mutate(owner=!!ownership_function,
         ID=!!PWSID_function,
         Municipality=!!water_system_function,
         Date = my(paste(Month,"-",Year))) %>%
         arrange(Date)

return(function_dataframe)
}
```

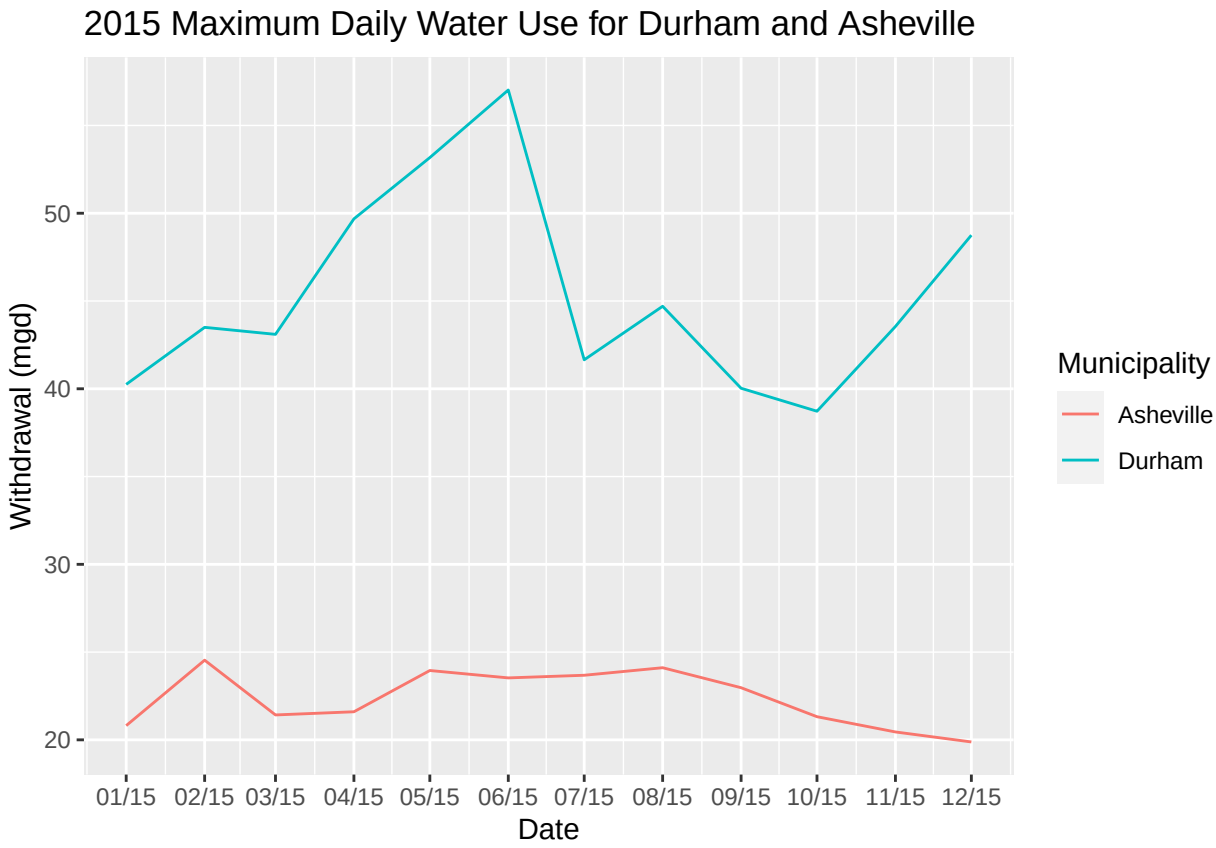
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID=‘03-32-010’) for each month in 2015

```
#7

the_year<- 2015
the_PWSID <-'03-32-010'

Durham_2015 <-water_function(the_year, the_PWSID)


ggplot(Durham_2015, aes(x=Date, y=Max.Daily.Use))+
  geom_line()+
  labs(title = paste("2015 Maximum Daily Water Use for Durham"),
       y="Withdrawal (mgd)",
       x="Date") +
  scale_x_date(date_labels="%m/%y", date_breaks="1 month")
```



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

the_year<- 2015
the_PWSID <-'01-11-010'

Asheville_2015 <-water_function(the_year, the_PWSID)

combined_df <- rbind(Asheville_2015, Durham_2015)

ggplot(combined_df, aes(x=Date, y=Max.Daily.Use, color=Municipality))+
  geom_line()+
  labs(title = paste("2015 Maximum Daily Water Use for Durham and Asheville"),
       y="Withdrawal (mgd)",
       x="Date") +
  scale_x_date(date_labels="%m/%y", date_breaks="1 month")
```

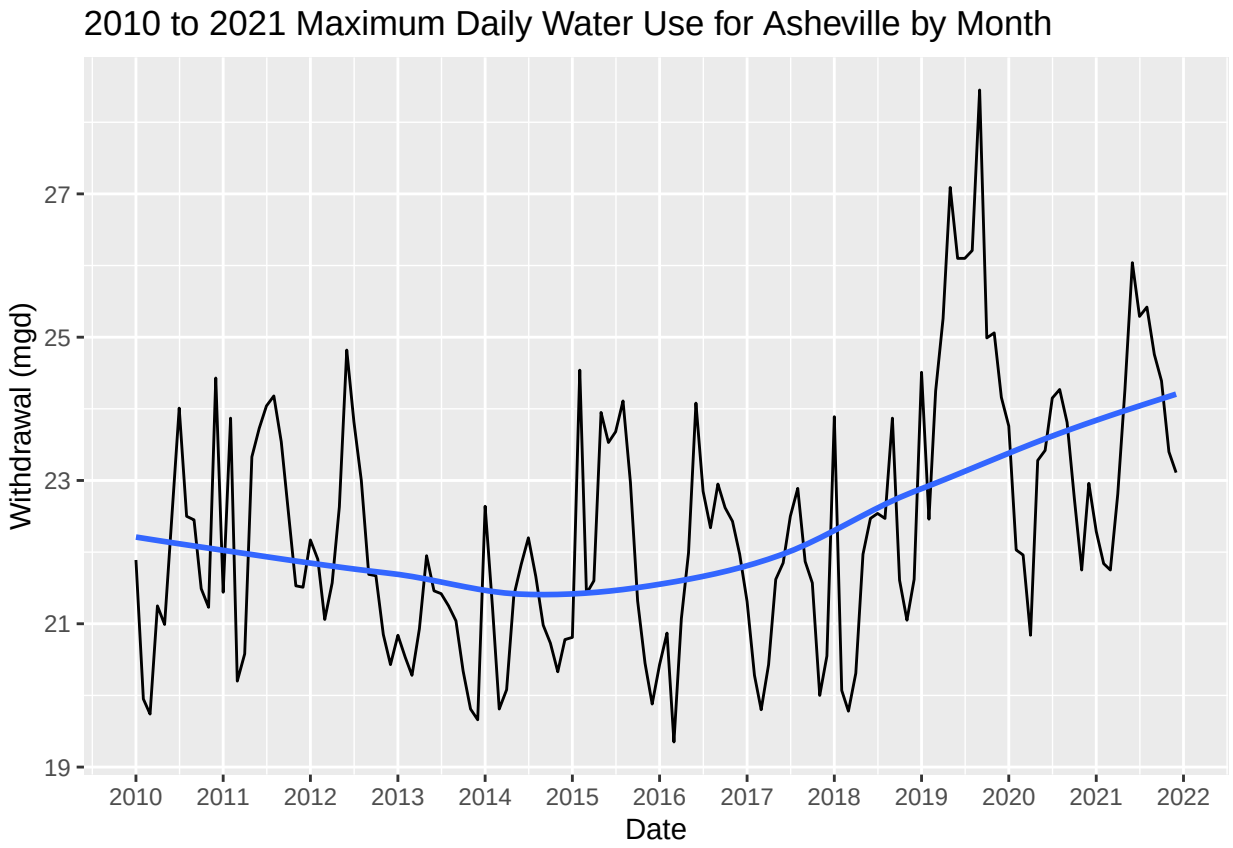


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2021.Add a smoothed line to the plot (method = 'loess').

   TIP: See Section 3.2 in the "10_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_year <- rep(2010:2021)
the_PWSID <- '01-11-010'

Asheville_years_df <- lapply(X=the_year, FUN=water_function, the_PWSID)

Asheville_years_df <- bind_rows(Asheville_years_df)

ggplot(Asheville_years_df, aes(x=Date, y=Max.Daily.Use))+
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = "2010 to 2021 Maximum Daily Water Use for Asheville by Month",
       y="Withdrawal (mgd)",
       x="Date")+
  scale_x_date(date_labels="%Y", date_breaks="1 year")
```

```
## `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? > Answer: Asheville's water use decreased from 2010 to 2014 or 2015 and then increased from 2014 or 2015 to 2021. >