Assignment 10: Data Scraping

Hannah Wudke

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(here)
library(dplyr)
library(rvest)
library(ggplot2)
library(ggridges)

here()
```

```
## [1] "/home/guest/EDE-Class"
```

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham's 2023 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=2023

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

```
#2
water.webpage <-
  read_html('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2023')
```

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

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

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

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

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

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

```
MGD <- water.webpage %>% html_nodes("th~ td+ td") %>% html_text()
MGD
```

```
##  [1] "28.9000" "33.3000" "43.7000" "30.0000" "40.0000" "37.2300" "34.2000"
##  [8] "44.9000" "40.3500" "30.9000" "56.7000" "33.3000"
```

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.)

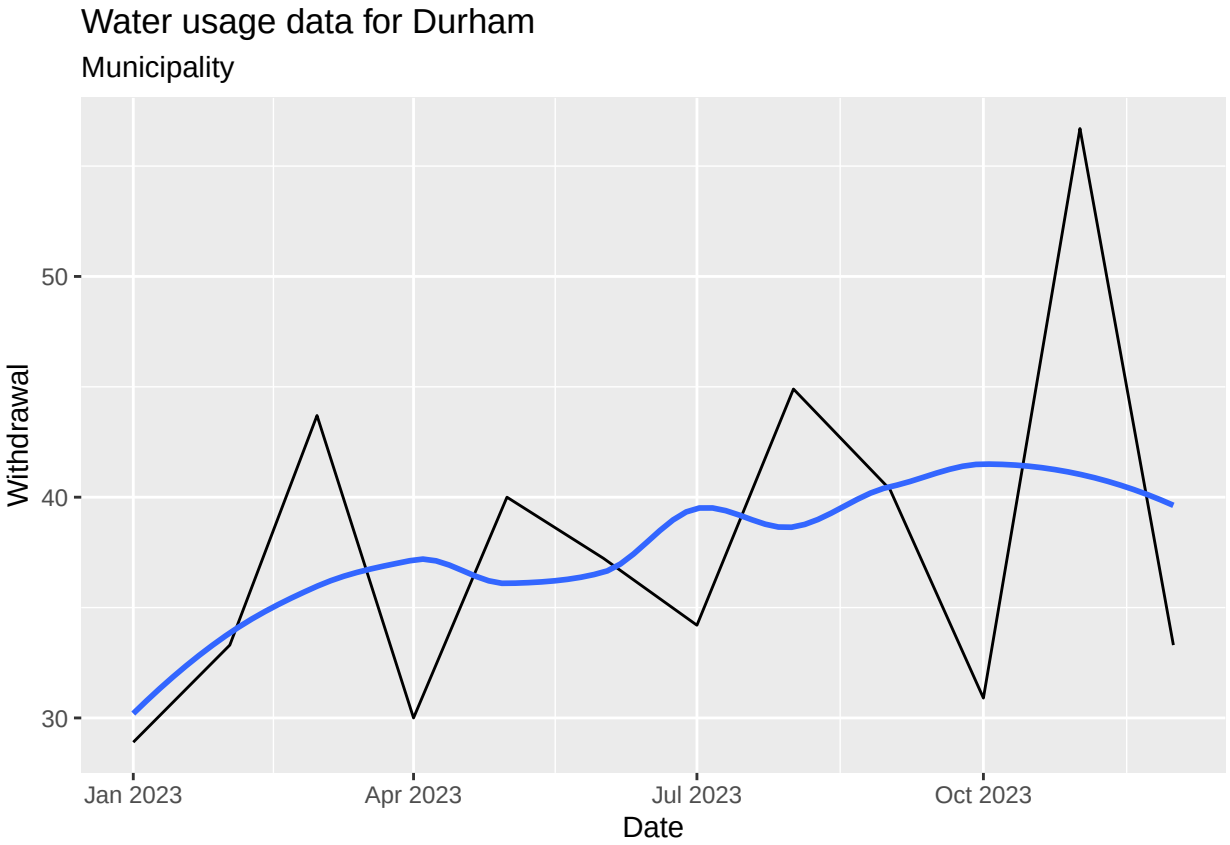
   TIP: Use `rep()` to repeat a value when creating a dataframe.

   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 2023, making sure, the months are presented in proper sequence.

```
#4
DurhamWater <- data.frame("Month" = rep(1:12), "Year" = rep(2023,12),
                          "Max.Day.Use" = as.numeric(MGD))

DurhamWater <- DurhamWater %>%
  mutate(Ownership = !!Ownership,
         PWSID = !!PWSID,
         Water.System.Name = !!Water.System.Name, Date = my(paste(Month, "-", Year)))
#5
Durham.Water.Plot <- ggplot(DurhamWater,
            aes(x=Date,y=Max.Day.Use)) + geom_line() + geom_smooth(method="loess",
            se=FALSE) + labs(title = paste("Water usage data for",
            Water.System.Name), subtitle = Ownership, y="Withdrawal", x="Date")

print(Durham.Water.Plot)
```

```
## `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, returning a dataframe. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

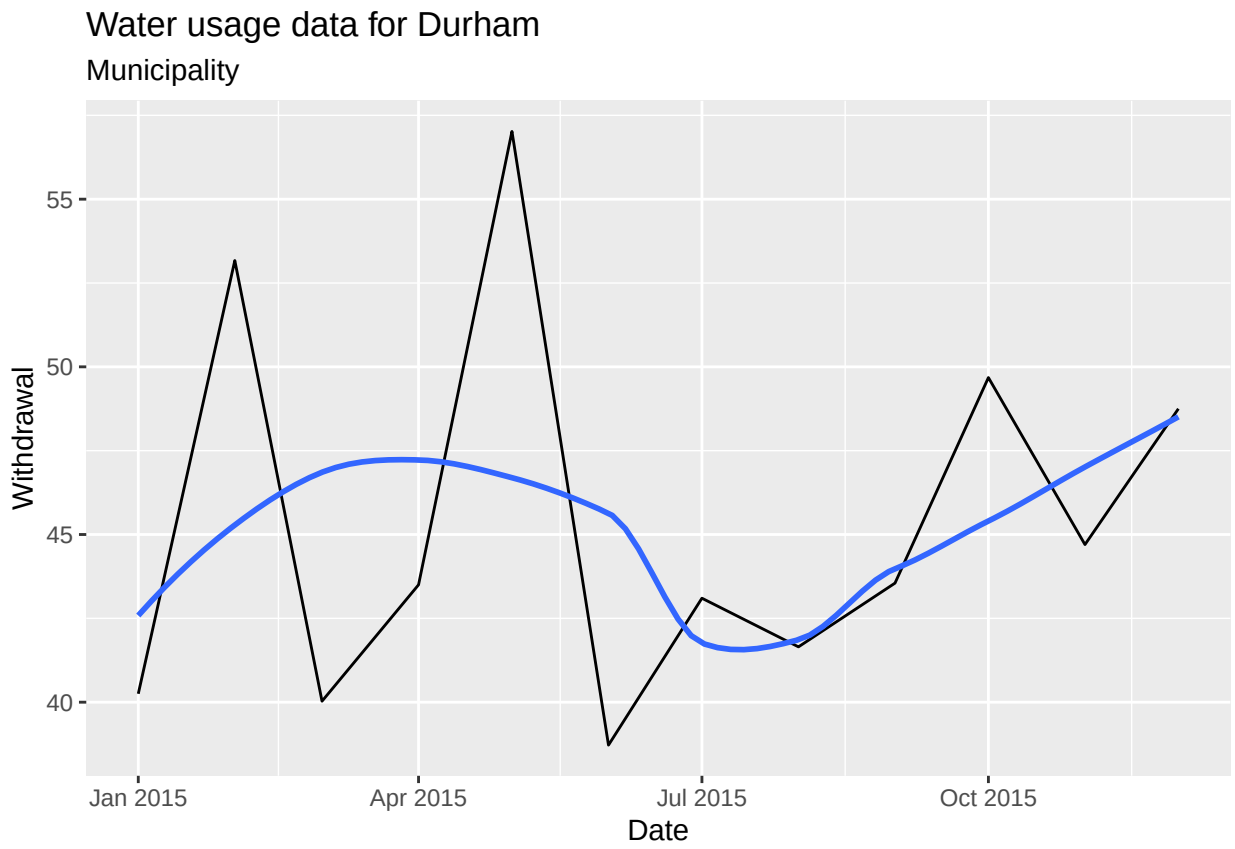
```
#6.
scrape.fxn <- function(Year, PWSID){
  Water.Website <- read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',
                                    PWSID, '&year=', Year))

Ownership.Tag <- 'div+ table tr:nth-child(2) td:nth-child(4)'
MGD.Tag <- 'th~ td+ td'
PWSID.Tag <- 'td tr:nth-child(1) td:nth-child(5)'
Water.System.Name.Tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'

The.Ownership <- Water.Website %>% html_nodes(Ownership.Tag) %>% html_text()
The.MGD <- Water.Website %>% html_nodes(MGD.Tag) %>% html_text()
The.PWSID <- Water.Website %>% html_nodes(PWSID.Tag) %>% html_text()
The.Water.System.Name <- Water.Website %>% html_nodes(Water.System.Name.Tag) %>% html_text()

The.Dataframe <- data.frame("Month" = rep(1:12), "Year" = rep(Year,12),
  "Average.Withdrawals.MGD" = as.numeric(The.MGD)) %>% mutate(Ownership = !!The.Ownership,
        Water.System.Name = !!The.Water.System.Name, PWSID = !!The.PWSID,
        Date = my(paste(Month,"-",Year)))
```

```
Sys.sleep(3)

return(The.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
Durham2015 <- scrape.fxn(2015, '03-32-010')

Durham.Water.Plot.2015 <- ggplot(Durham2015,
            aes(x=Date,
          y=Average.Withdrawals.MGD)) + geom_line() + geom_smooth(method="loess",
            se=FALSE) + labs(title = paste("Water usage data for",
            Water.System.Name), subtitle = Ownership, y="Withdrawal", x="Date")
print(Durham.Water.Plot.2015)
```

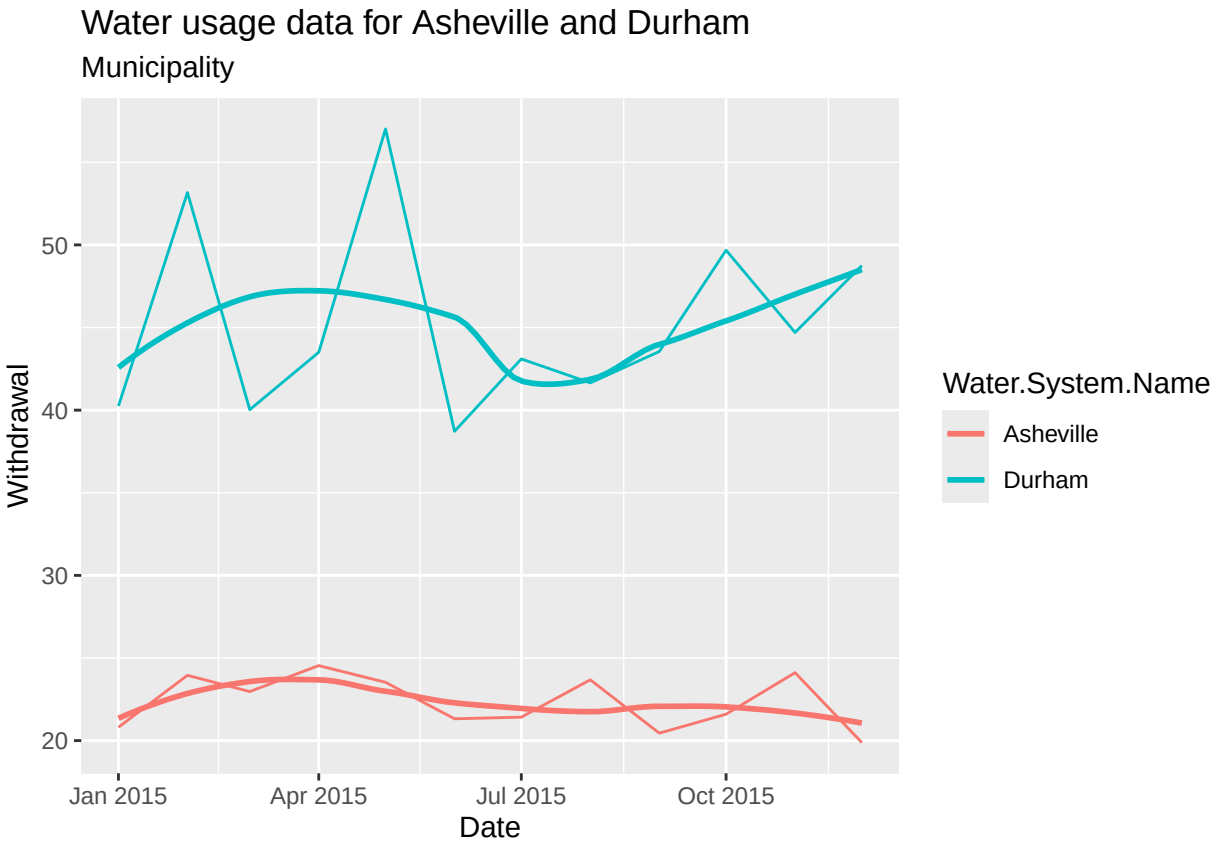
```
## `geom_smooth()` using formula = 'y ~ x'
```



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.

```r
#8
Asheville2015 <- scrape.fxn(2015, '01-11-010')
view(Asheville2015)

Asheville.Durham <- bind_rows(Asheville2015, Durham2015)

 Asheville.Durham.Plot <- ggplot(Asheville.Durham,
            aes(x=Date,y=Average.Withdrawals.MGD,
            color=Water.System.Name)) + geom_line() + geom_smooth(method="loess",
            se=FALSE) + labs(title = paste("Water usage data for Asheville and Durham"),
                             subtitle = Ownership, y="Withdrawal", x="Date")
print(Asheville.Durham.Plot)
```

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

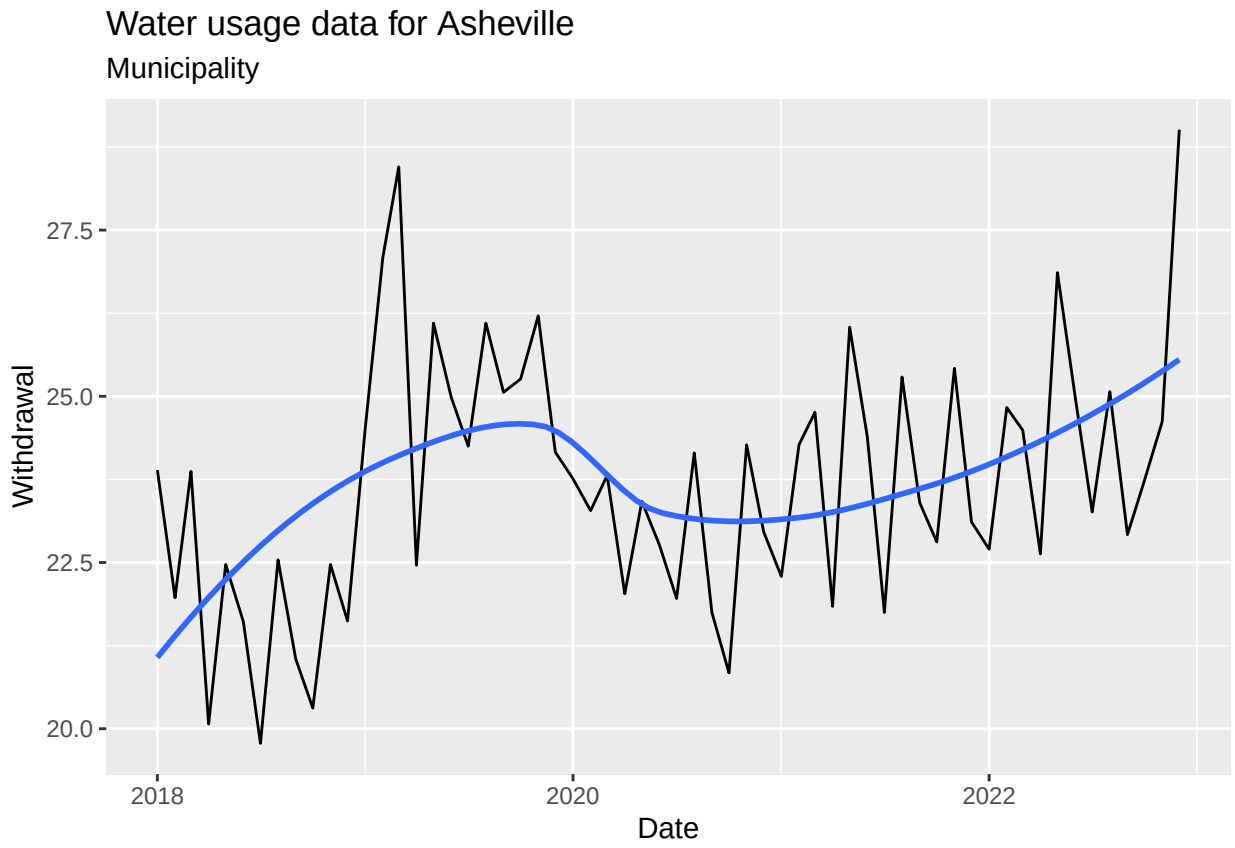


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2018 thru 2022.Add a smoothed line to the plot (method = 'loess').

   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
Asheville2018 <- scrape.fxn(2018, '01-11-010')
Asheville2019 <- scrape.fxn(2019, '01-11-010')
Asheville2020 <- scrape.fxn(2020, '01-11-010')
Asheville2021 <- scrape.fxn(2021, '01-11-010')
Asheville2022 <- scrape.fxn(2022, '01-11-010')

Asheville.Years <- bind_rows(Asheville2018, Asheville2019, Asheville2020,
                             Asheville2021, Asheville2022)
 Asheville.Years.Plot <- ggplot(Asheville.Years,
  aes(x=Date,y=Average.Withdrawals.MGD)) + geom_line() + geom_smooth(method="loess",
            se=FALSE) + labs(title = paste("Water usage data for Asheville"),
                             subtitle = Ownership, y="Withdrawal", x="Date")
print(Asheville.Years.Plot)
```

```
## `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: Yes, water usage increases from 2018 to 2020, decreases for a year, likely due to the pandemic, and then steadily increases again towards 2022. >