Assignment 10: Data Scraping

Lucy Wang

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(viridis);library(here)
library(rvest)
here()
```

```
## [1] "/Users/lucywang/Documents/EDE_Fall2023"
```

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
theURL <- 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
the_name <- theURL %>%
  html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>%
  html_text()
PWSID <- theURL %>%
  html_nodes('td tr:nth-child(1) td:nth-child(5)') %>%
  html_text()
the_ownership <- theURL %>%
  html_nodes('div+ table tr:nth-child(2) td:nth-child(4)') %>%
  html_text()
max_day <- theURL %>%
  html_nodes('th~ td+ td , th~ td+ td') %>%
  html_text()
```

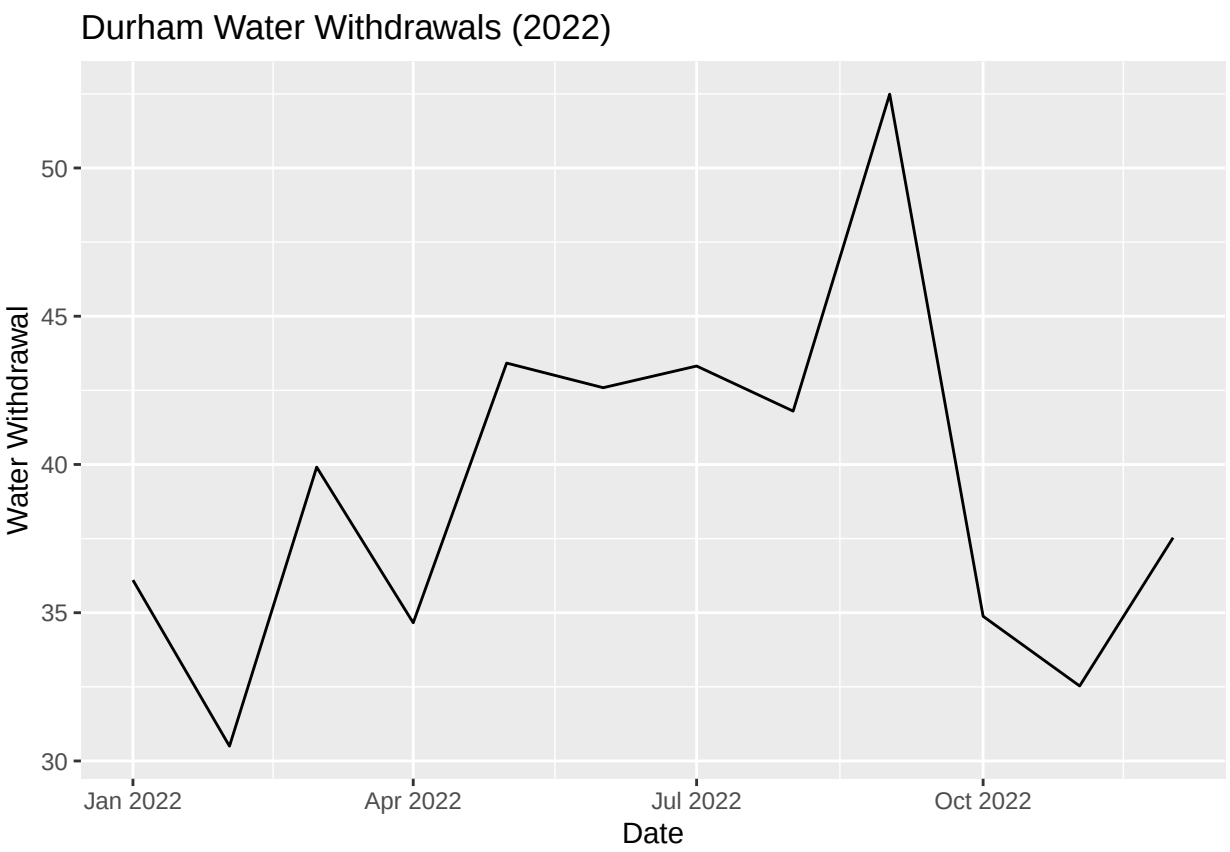
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   > 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 maximum daily withdrawals across the months for 2022

```r
#4
the_months <- theURL %>%
  html_nodes('.fancy-table:nth-child(31) tr+ tr th') %>%
  html_text()

LWSP_df <- data.frame(
  'Month'= month.name[match(the_months,month.abb)],
  'Water_System_Name' = rep(the_name,12),
  'PWSID' = rep(PWSID,12),
  'Ownership' = rep(the_ownership,12),
  'Max_Day_Use' = as.numeric(max_day), stringsAsFactors = T
) %>%
  mutate(Date = my(paste(Month,'-',2022))) %>%
  arrange(Date)

#5
ggplot(LWSP_df, aes(x=Date, y=Max_Day_Use))+
  geom_line()+
  labs(title = 'Durham Water Withdrawals (2022)',
   y="Water Withdrawal",
   x="Date")
```

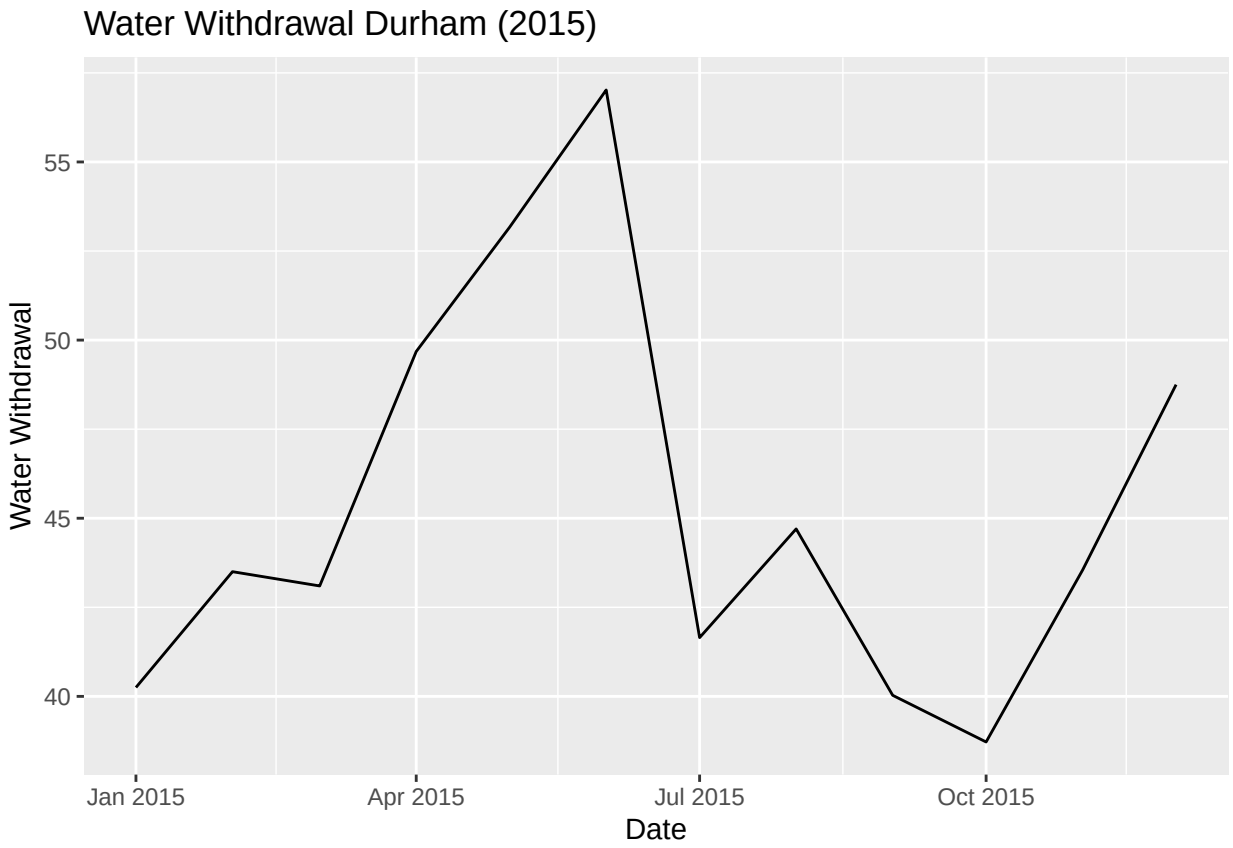


6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

```r
#6.
scrape_data <- function(the_PWSID, the_year){
  the_url <- paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=', the_PWSID, '&year=', the_
  the_website <- read_html(the_url)

  the_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()
  the_ownership <- the_website %>%
    html_nodes('div+ table tr:nth-child(2) td:nth-child(4)') %>%
    html_text()
  max_day <- the_website %>%
    html_nodes('th~ td+ td , th~ td+ td') %>%
    html_text()
  the_months <- c('Jan', 'May', 'Sep',
                  'Feb', 'Jun', 'Oct',
                  'Mar', 'Jul', 'Nov',
                  'Apr', 'Aug', 'Dec')

  the_df <- data.frame(
    'Month'= month.name[match(the_months,month.abb)],
    'Water_System_Name' = rep(the_name,12),
    'PWSID' = rep(PWSID,12),
    'Ownership' = rep(the_ownership,12),
    'Max_Day_Use' = as.numeric(max_day), stringsAsFactors = T
  ) %>%
    mutate(Date = my(paste(Month,'-',the_year))) %>%
    arrange(Date)

  return(the_df)
}
```

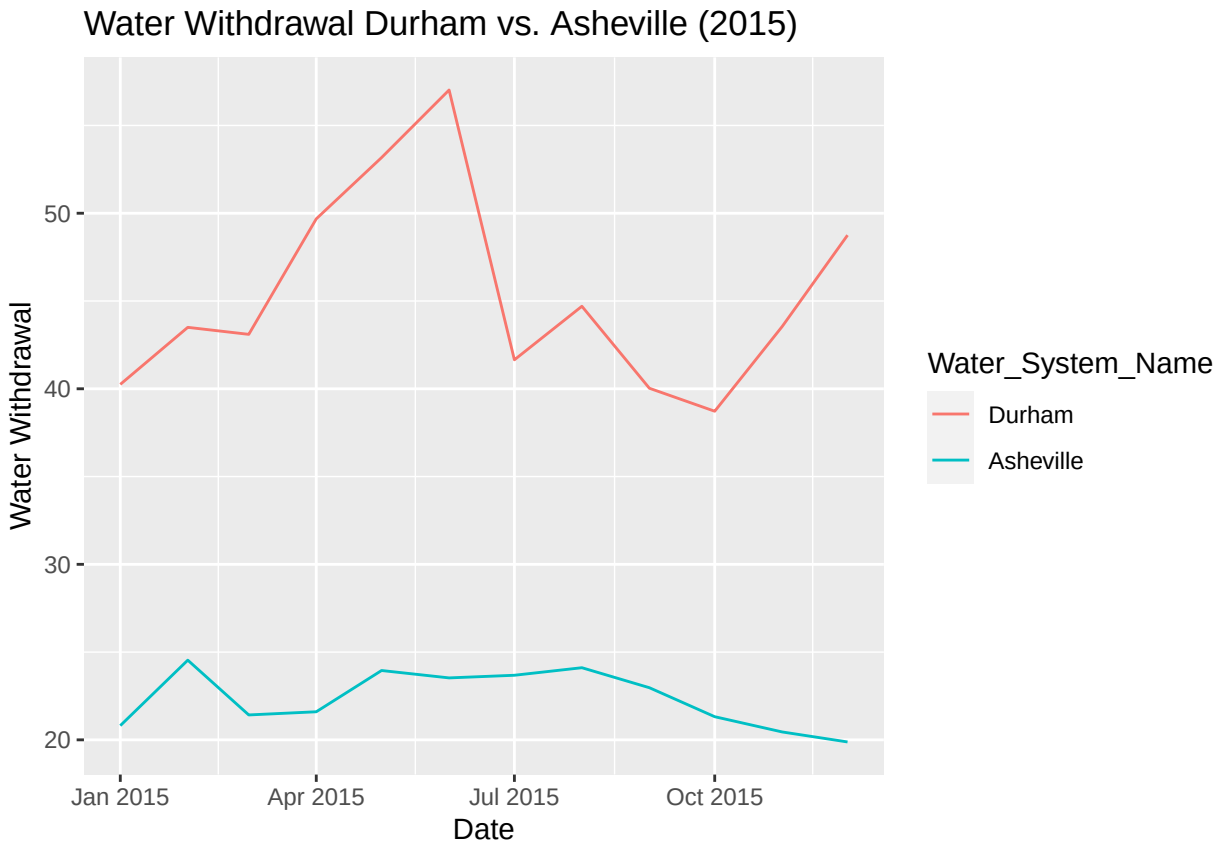
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID=‘03-32-010’) for each month in 2015

```r
#7
Durham_2015 <- scrape_data('03-32-010',2015)

ggplot(Durham_2015, aes(x=Date, y=Max_Day_Use))+
  geom_line() +
  labs(title = 'Water Withdrawal Durham (2015)',
   y="Water Withdrawal",
   x="Date")
```



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
Asheville_2015 <- scrape_data('01-11-010', 2015)
Durham_Asheville_2015 <- rbind(Durham_2015, Asheville_2015)

ggplot(Durham_Asheville_2015, aes(x=Date,y=Max_Day_Use,color=Water_System_Name))+
  geom_line()+
    labs(title = 'Water Withdrawal Durham vs. Asheville (2015)',
     y="Water Withdrawal",
     x="Date")
```

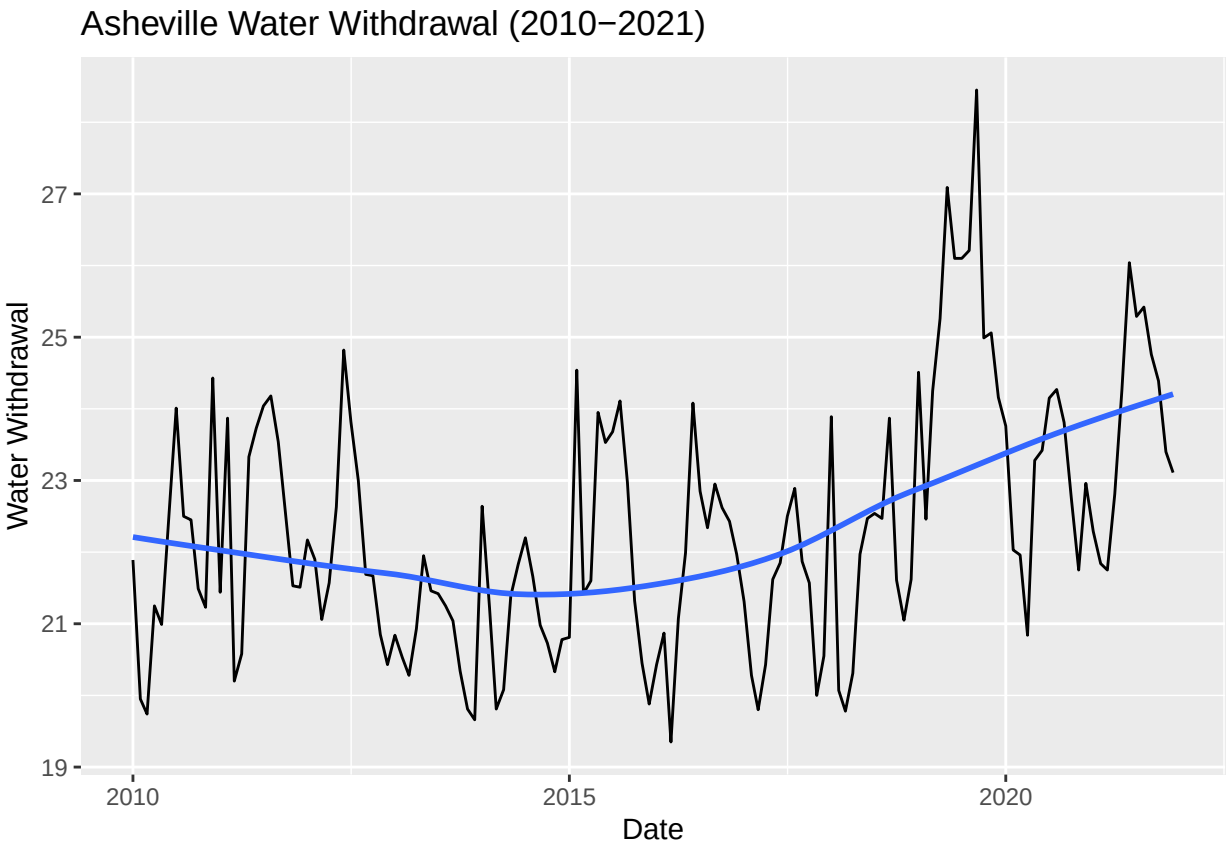


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2021.Add a smoothed line to the plot (method = 'loess').

   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_years <- c(2010:2021)
Asheville_PWSID <- '01-11-010'

the_dfs <- lapply(the_years, function(year) scrape_data(Asheville_PWSID, the_year = year))

Asheville_df <- bind_rows(the_dfs)

ggplot(Asheville_df, aes(x=Date, y=Max_Day_Use))+
  geom_line()+
  geom_smooth(method='loess', se=FALSE)+
  labs(title = 'Asheville Water Withdrawal (2010-2021)',
     y="Water Withdrawal",
     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? > Answer: > It started with a slight decrease in the early 2010s but started increasing since 2015.