Assignment 10: Data Scraping

Lauren Ng

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(rvest)
library(dplyr)
library(ggplot2)
library(lubridate)
```

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
webpage <- read_html(
  'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2023')
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
- 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
watersystemname <- webpage %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()
watersystemname
```

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

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

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

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

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

```
maxdayuse <- webpage %>%
  html_nodes("th~ td+ td") %>%
  html_text()
maxdayuse
```

```
##  [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"
```

```
month <- webpage %>%
  html_nodes(".fancy-table:nth-child(31) tr+ tr th")%>%
html_text()
month
```

```
##  [1] "Jan" "May" "Sep" "Feb" "Jun" "Oct" "Mar" "Jul" "Nov" "Apr" "Aug" "Dec"
```

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.)

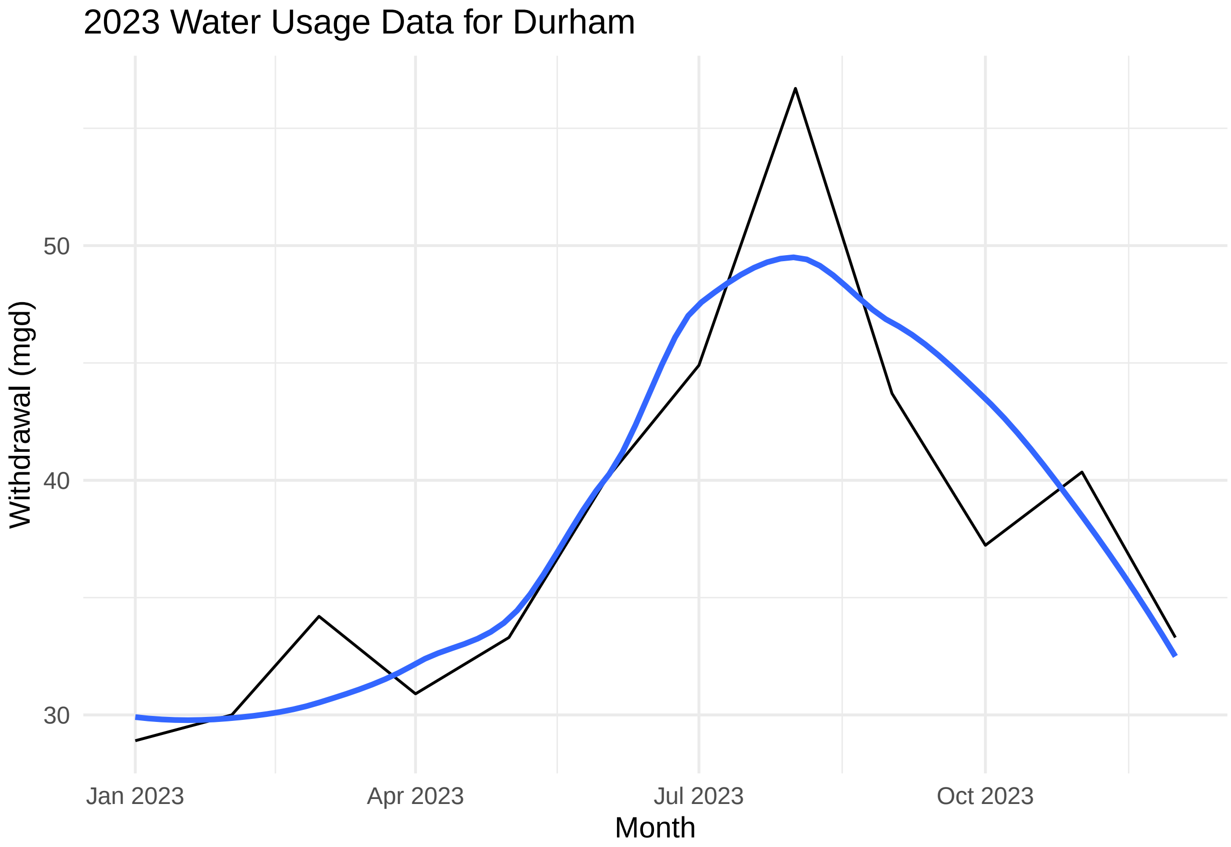
   TIP: Use `rep()` to repeat a value when creating a dataframe.

   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

df_withdrawals <- data.frame(
  "Month" = c(1,5,9,2,6,10,3,7,11,4,8,12),
  "Year" = rep(2023, 12),
  "Water System" = watersystemname,
  "PWSID" = PWSID,
  "Ownership" = ownership,
  "Max_Daily_Use" = as.numeric(maxdayuse)
) %>%
mutate(Date = make_date(Year, Month, 1))

#5
ggplot(df_withdrawals, aes(x = Date, y = `Max_Daily_Use`)) +
  geom_line() +
  geom_smooth(method = "loess", se = FALSE) +
  labs(
    title = paste("2023 Water Usage Data for", watersystemname),
    y = "Withdrawal (mgd)",
    x = "Month"
  ) +
  theme_minimal()
```

```
## ‘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 func
**Be sure to modify the code to reflect the year and site (pwsid) scraped**.


‘‘‘ r
#6
scrape.it <- function(pwsid,the_year){

the_scrape_url <-  read_html(
paste0(
’https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=’, pwsid,’&’, ’year=’,the_year)
)


#Set the element address variables (determined in the previous step)

watersystemname_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)’
maxdayuse_tag <- ’th~ td+ td’

#Scrape the data items
watersystemname <- the_scrape_url %>%
html_nodes(watersystemname_tag) %>%
html_text()
PWSID <- the_scrape_url %>%
html_nodes(PWSID_tag) %>%
html_text()
ownership <- the_scrape_url %>%
html_nodes(ownership_tag) %>%
html_text()
maxdayuse <- the_scrape_url %>%
html_nodes(maxdayuse_tag) %>%
html_text()

#Convert to a dataframe
df_withdrawals <- data.frame(
  "Month" = c(1,5,9,2,6,10,3,7,11,4,8,12),
  "Year" = rep(the_year, 12),
  "Water System" = watersystemname,
  "PWSID" = PWSID,
  "Ownership" = ownership,
  "Max_Daily_Use" = as.numeric(maxdayuse)
) %>%
mutate(Date = make_date(Year, Month, 1))

return(df_withdrawals)
}
```

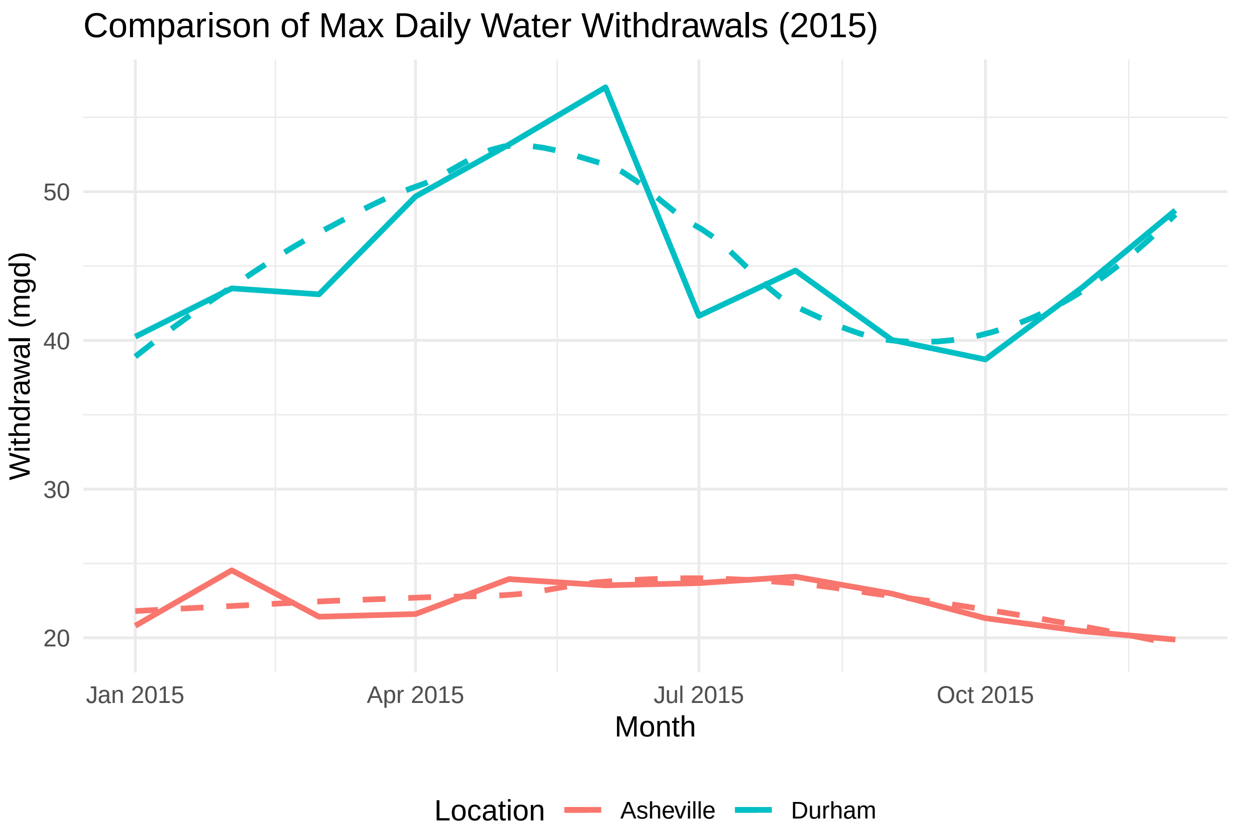
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID=‘03-32-010’) for each month in 2015

```
#7
Durham_2015 <- scrape.it('03-32-010',2015)
view(Durham_2015)
```

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.it('01-11-010',2015)
view(Asheville_2015)

combined_data <- bind_rows(
  Durham_2015 %>% mutate(Location = "Durham"),
  Asheville_2015 %>% mutate(Location = "Asheville")
)

ggplot(combined_data, aes(x = Date, y = Max_Daily_Use, color = Location)) +
  geom_line(size = 1) +
  geom_smooth(method = "loess", se = FALSE, linetype = "dashed") +
  labs(
    title = "Comparison of Max Daily Water Withdrawals (2015)",
    y = "Withdrawal (mgd)",
    x = "Month",
    color = "Location"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.

## ‘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.

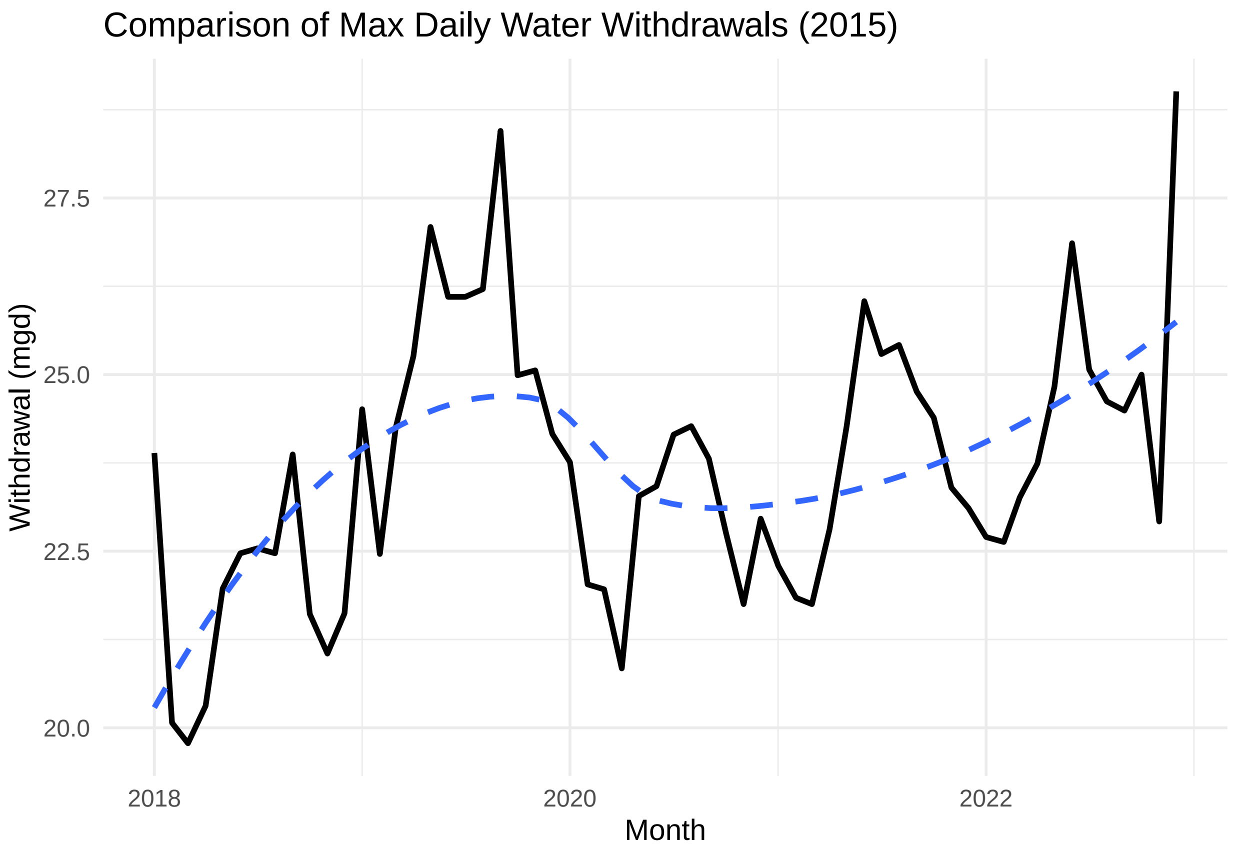
```
#9
asheville_pwsid <- "01-11-010"
years <- 2018:2022

# Scrape data for Asheville for the specified years using map2()
```

```r
#asheville_data <- map2(asheville_pwsid,years, scrape.it)%>%
asheville_data <- map2_dfr(rep(asheville_pwsid, length(years)), years, scrape.it)
bind_rows(asheville_data)
```

```
##    Month Year Water.System     PWSID    Ownership Max_Daily_Use       Date
## 1      1 2018    Asheville 01-11-010 Municipality         23.89 2018-01-01
## 2      5 2018    Asheville 01-11-010 Municipality         21.97 2018-05-01
## 3      9 2018    Asheville 01-11-010 Municipality         23.87 2018-09-01
## 4      2 2018    Asheville 01-11-010 Municipality         20.07 2018-02-01
## 5      6 2018    Asheville 01-11-010 Municipality         22.47 2018-06-01
## 6     10 2018    Asheville 01-11-010 Municipality         21.61 2018-10-01
## 7      3 2018    Asheville 01-11-010 Municipality         19.78 2018-03-01
## 8      7 2018    Asheville 01-11-010 Municipality         22.54 2018-07-01
## 9     11 2018    Asheville 01-11-010 Municipality         21.05 2018-11-01
## 10     4 2018    Asheville 01-11-010 Municipality         20.31 2018-04-01
## 11     8 2018    Asheville 01-11-010 Municipality         22.47 2018-08-01
## 12    12 2018    Asheville 01-11-010 Municipality         21.62 2018-12-01
## 13     1 2019    Asheville 01-11-010 Municipality         24.51 2019-01-01
## 14     5 2019    Asheville 01-11-010 Municipality         27.09 2019-05-01
## 15     9 2019    Asheville 01-11-010 Municipality         28.45 2019-09-01
## 16     2 2019    Asheville 01-11-010 Municipality         22.46 2019-02-01
## 17     6 2019    Asheville 01-11-010 Municipality         26.10 2019-06-01
## 18    10 2019    Asheville 01-11-010 Municipality         24.99 2019-10-01
## 19     3 2019    Asheville 01-11-010 Municipality         24.25 2019-03-01
## 20     7 2019    Asheville 01-11-010 Municipality         26.10 2019-07-01
## 21    11 2019    Asheville 01-11-010 Municipality         25.06 2019-11-01
## 22     4 2019    Asheville 01-11-010 Municipality         25.26 2019-04-01
## 23     8 2019    Asheville 01-11-010 Municipality         26.21 2019-08-01
## 24    12 2019    Asheville 01-11-010 Municipality         24.16 2019-12-01
## 25     1 2020    Asheville 01-11-010 Municipality         23.76 2020-01-01
## 26     5 2020    Asheville 01-11-010 Municipality         23.28 2020-05-01
## 27     9 2020    Asheville 01-11-010 Municipality         23.81 2020-09-01
## 28     2 2020    Asheville 01-11-010 Municipality         22.03 2020-02-01
## 29     6 2020    Asheville 01-11-010 Municipality         23.42 2020-06-01
## 30    10 2020    Asheville 01-11-010 Municipality         22.76 2020-10-01
## 31     3 2020    Asheville 01-11-010 Municipality         21.96 2020-03-01
## 32     7 2020    Asheville 01-11-010 Municipality         24.15 2020-07-01
## 33    11 2020    Asheville 01-11-010 Municipality         21.75 2020-11-01
## 34     4 2020    Asheville 01-11-010 Municipality         20.84 2020-04-01
## 35     8 2020    Asheville 01-11-010 Municipality         24.27 2020-08-01
## 36    12 2020    Asheville 01-11-010 Municipality         22.96 2020-12-01
## 37     1 2021    Asheville 01-11-010 Municipality         22.29 2021-01-01
## 38     5 2021    Asheville 01-11-010 Municipality         24.27 2021-05-01
## 39     9 2021    Asheville 01-11-010 Municipality         24.76 2021-09-01
## 40     2 2021    Asheville 01-11-010 Municipality         21.84 2021-02-01
## 41     6 2021    Asheville 01-11-010 Municipality         26.04 2021-06-01
## 42    10 2021    Asheville 01-11-010 Municipality         24.39 2021-10-01
## 43     3 2021    Asheville 01-11-010 Municipality         21.75 2021-03-01
## 44     7 2021    Asheville 01-11-010 Municipality         25.29 2021-07-01
## 45    11 2021    Asheville 01-11-010 Municipality         23.40 2021-11-01
## 46     4 2021    Asheville 01-11-010 Municipality         22.81 2021-04-01
## 47     8 2021    Asheville 01-11-010 Municipality         25.42 2021-08-01
## 48    12 2021    Asheville 01-11-010 Municipality         23.11 2021-12-01
## 49     1 2022    Asheville 01-11-010 Municipality         22.70 2022-01-01
## 50     5 2022    Asheville 01-11-010 Municipality         24.83 2022-05-01
## 51     9 2022    Asheville 01-11-010 Municipality         24.49 2022-09-01
## 52     2 2022    Asheville 01-11-010 Municipality         22.63 2022-02-01
## 53     6 2022    Asheville 01-11-010 Municipality         26.86 2022-06-01
## 54    10 2022    Asheville 01-11-010 Municipality         25.00 2022-10-01
## 55     3 2022    Asheville 01-11-010 Municipality         23.26 2022-03-01
## 56     7 2022    Asheville 01-11-010 Municipality         25.07 2022-07-01
## 57    11 2022    Asheville 01-11-010 Municipality         22.92 2022-11-01
## 58     4 2022    Asheville 01-11-010 Municipality         23.74 2022-04-01
## 59     8 2022    Asheville 01-11-010 Municipality         24.62 2022-08-01
## 60    12 2022    Asheville 01-11-010 Municipality         29.01 2022-12-01
```

```
ggplot(asheville_data, aes(x = Date, y = Max_Daily_Use)) +
  geom_line(size = 1) +
  geom_smooth(method = "loess", se = FALSE, linetype = "dashed") +
  labs(
    title = "Comparison of Max Daily Water Withdrawals (2015)",
    y = "Withdrawal (mgd)",
    x = "Month"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```

```
## `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, looking at the plot, it appears that Asheville does have a trend in increasing water usage over time. >