Assignment 10: Data Scraping

Jenn McNeill

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

```r
#1
library(tidyverse)
library(rvest)

getwd()
```

```
## [1] "/Users/jennifermcneill/EDE_Fall2023/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
#bring in the website with the whole web address
the_website <- 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
#set the element address variables
Water_System_Name_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)'
Maximum_Day_Use_MGD_Tag <- 'th~ td+ td'
```

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.)

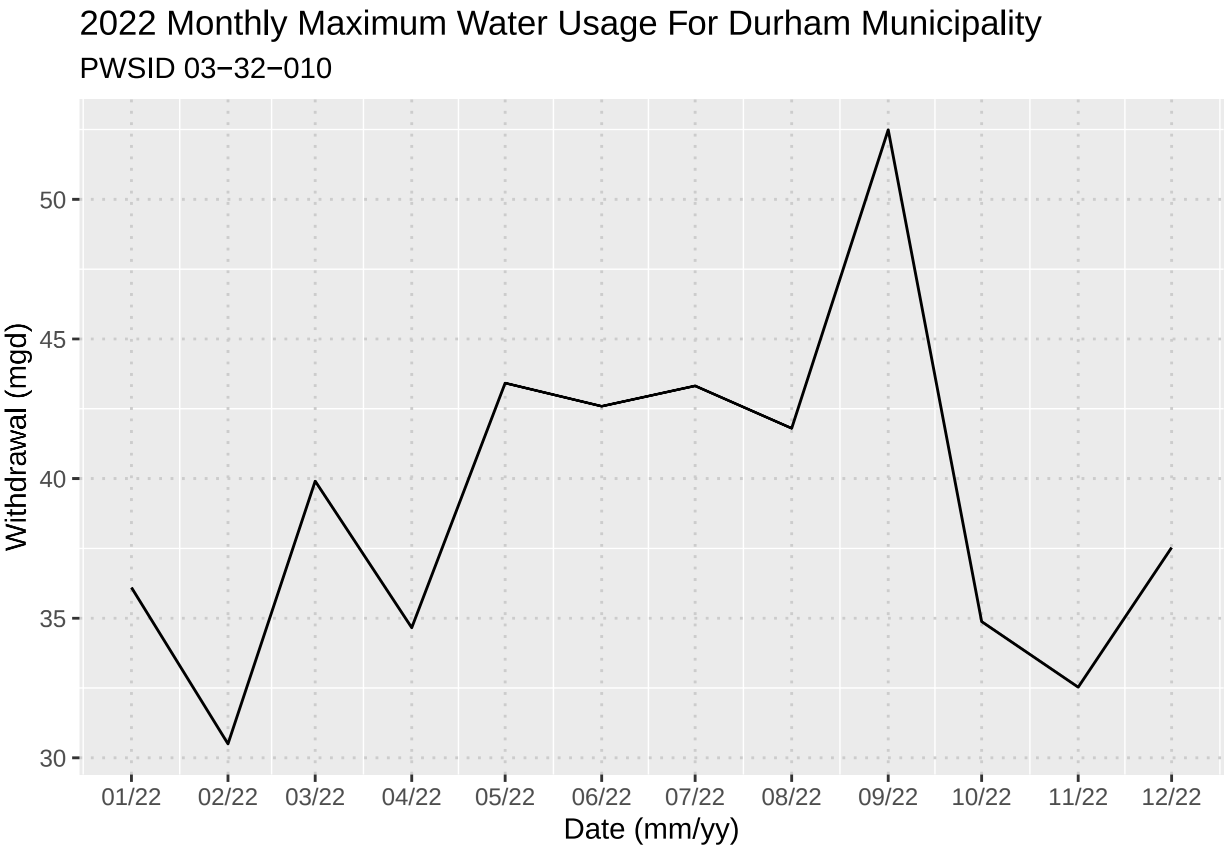
   > TIP: Use `rep()` to repeat a value when creating a dataframe.

   > 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
#construct the scraping web address for later
the_base_url <- 'https://www.ncwater.org/WUDC/app/LWSP/report.php'
the_url_pwsid <- '03-32-010'
the_url_year <- 2022
the_scrape_url <- paste0(the_base_url, '?pwsid=', the_url_pwsid, '&year=', the_url_year)
the_website <- read_html(the_scrape_url)
```

```r
#scrape the data items from the tags created above
Water_System_Name <- the_website %>% html_nodes(Water_System_Name_Tag) %>% html_text()
PWSID <- the_website %>% html_nodes(PWSID_Tag) %>%  html_text()
Ownership <- the_website %>% html_nodes(Ownership_Tag) %>% html_text()
Maximum_Day_Use_MGD <- the_website %>% html_nodes(Maximum_Day_Use_MGD_Tag) %>% html_text()

#convert to a dataframe
df_monthly_withdrawal <- data.frame(
  'Name' = rep(Water_System_Name,12),
  'PWSID' = rep(PWSID,12),
  'Ownership' = rep(Ownership,12),
  'Maximum_Day' = as.numeric(gsub(',','',Maximum_Day_Use_MGD)),
  'Year' = rep(the_url_year,12),
  'Month' = c("Jan","May","Sep","Feb","Jun","Oct",
              "Mar","Jul","Nov","Apr","Aug","Dec")) %>%
  mutate(Date = my(paste(Month,"-",Year)))


#5
#plot the Durham 2022 data
ggplot(df_monthly_withdrawal,
       aes(x = Date, y = Maximum_Day)) +
  geom_line() +
  labs(title = paste(the_url_year, "Monthly Maximum Water Usage For",
                     Water_System_Name, Ownership),
       subtitle = paste("PWSID", PWSID),
       y = "Withdrawal (mgd)",
       x = "Date (mm/yy)") +
  scale_x_date(date_labels = "%m/%y", date_breaks = "1 month") +
  theme(panel.grid.major = element_line(color = "grey80", linetype = "dotted"))
```



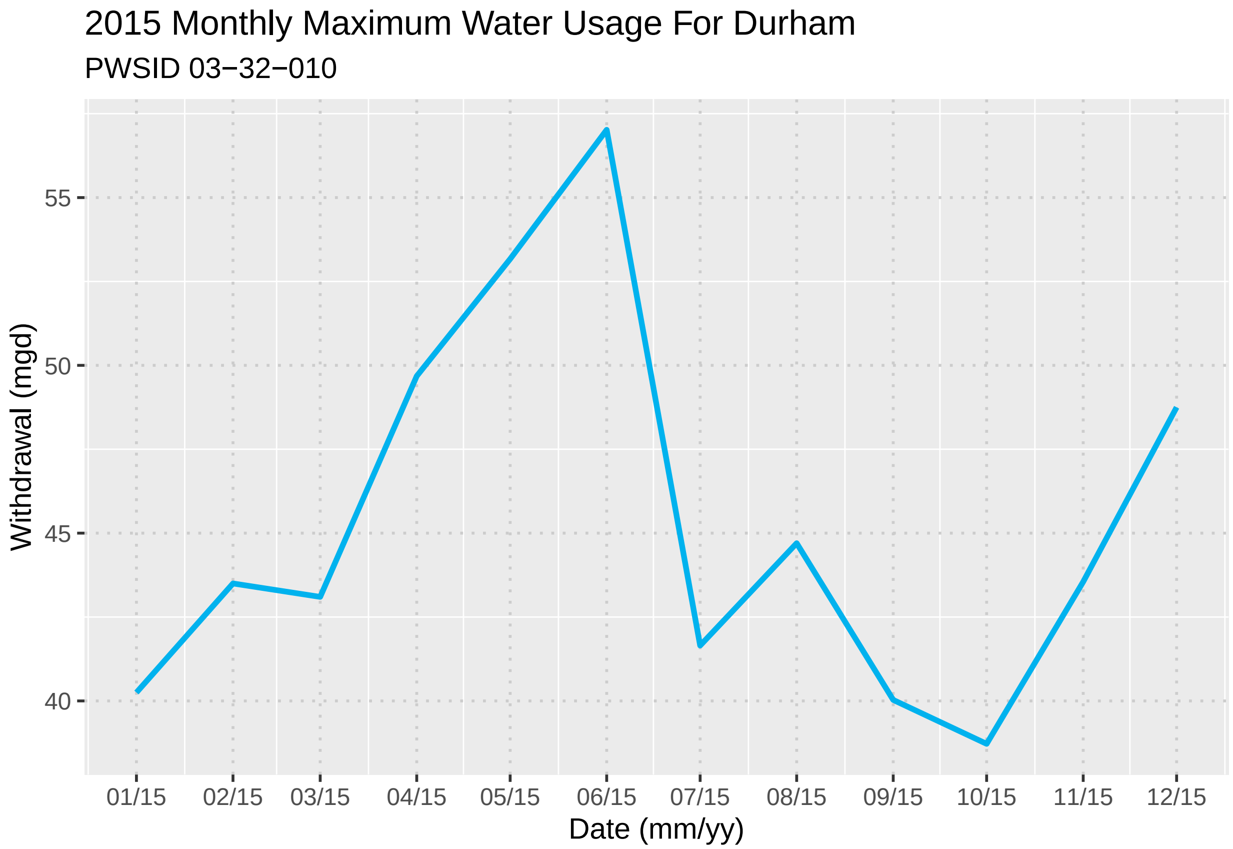
6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

```
#6.
#make a function that is dependent on year and pwsid
scraping_function <- function(the_url_year, the_url_pwsid){

  #retrieve the website contents
  the_website <- read_html(paste0(the_base_url, '?pwsid=',
                                  the_url_pwsid, '&year=', the_url_year))

  #set the element address variables (determined in the previous step)
  Water_System_Name_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)'
  Maximum_Day_Use_MGD_Tag <- 'th~ td+ td'

  #scrape the data items
  Water_System_Name <- the_website %>% html_nodes(Water_System_Name_Tag) %>% html_text()
  PWSID <- the_website %>% html_nodes(PWSID_Tag) %>%  html_text()
  Ownership <- the_website %>% html_nodes(Ownership_Tag) %>% html_text()
  Maximum_Day_Use_MGD <- the_website %>% html_nodes(Maximum_Day_Use_MGD_Tag) %>% html_text()

  #convert to a dataframe
  df_monthly_withdrawal <- data.frame(
    'Name' = rep(Water_System_Name,12),
    'PWSID' = rep(PWSID,12),
    'Ownership' = rep(Ownership,12),
    'Maximum_Day' = as.numeric(gsub(',','',Maximum_Day_Use_MGD)),
    'Year' = rep(the_url_year,12),
    'Month' = c("Jan","May","Sep","Feb","Jun","Oct",
              "Mar","Jul","Nov","Apr","Aug","Dec")) %>%
    mutate(Date = my(paste(Month,"-",Year)))

  #uncomment this if you are doing bulk scraping!
  #Sys.sleep(1)

  #return the dataframe
  return(df_monthly_withdrawal)
}
```

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

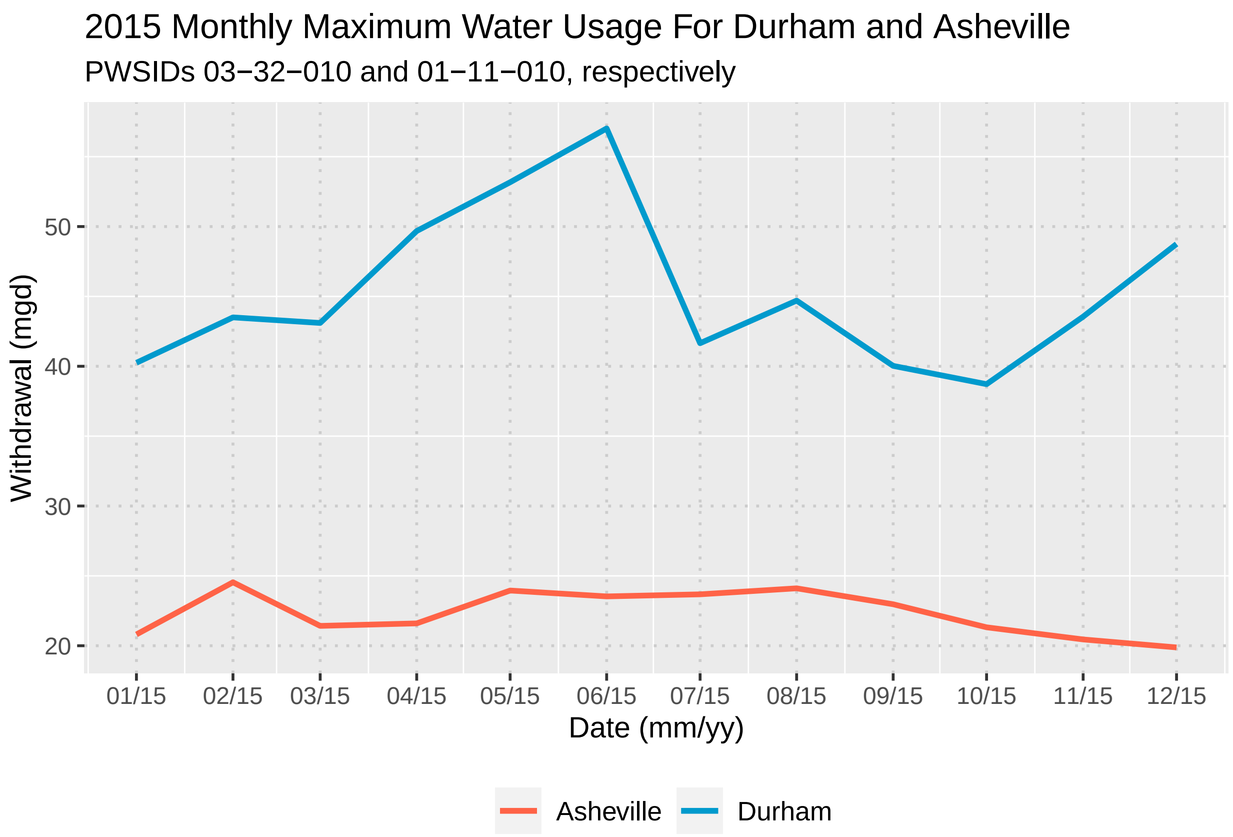
```
#7
#apply the scraping_function to Durham in 2015
Durham_2015 <- scraping_function(2015,'03-32-010')
view(Durham_2015)

#plot the Durham 2015 data
ggplot(Durham_2015, aes(x = Date)) +
  geom_line(aes(y = Maximum_Day), color = "deepskyblue2", size = 1) +
  labs(title = "2015 Monthly Maximum Water Usage For Durham",
       subtitle = "PWSID 03-32-010",
       y = "Withdrawal (mgd)",
       x = "Date (mm/yy)") +
  scale_x_date(date_labels = "%m/%y", date_breaks = "1 month") +
  theme(panel.grid.major = element_line(color = "grey80", linetype = "dotted"))
```

```
## 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.
```



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
#apply the scraping_function to Asheville in 2015
Asheville_2015 <- scraping_function(2015,'01-11-010')
view(Asheville_2015)

#merge Asheville 2015 data and Durham 2015 data into a single dataframe
Durham_vs_Asheville_2015 <- merge(Durham_2015, Asheville_2015, by = 'Date')
#note the column names of the Durham and Asheville data
view(Durham_vs_Asheville_2015)

#plot the Durham and Asheville 2015 data using the column names
ggplot(Durham_vs_Asheville_2015, aes(x = Date)) +
  geom_line(aes(y = Maximum_Day.x, color = "Durham"), size = 1) +
  geom_line(aes(y = Maximum_Day.y, color = "Asheville"), size = 1) +
  labs(title = "2015 Monthly Maximum Water Usage For Durham and Asheville",
       subtitle = "PWSIDs 03-32-010 and 01-11-010, respectively",
       y = "Withdrawal (mgd)",
       x = "Date (mm/yy)") +
  scale_x_date(date_labels = "%m/%y", date_breaks = "1 month") +
  scale_color_manual(values = c("Durham" = "deepskyblue3", "Asheville" = "tomato1"),
                     name = "") +
  theme(legend.position = "bottom", legend.text = element_text(size = 10),
        panel.grid.major = element_line(color = "grey80", linetype = "dotted"))
```

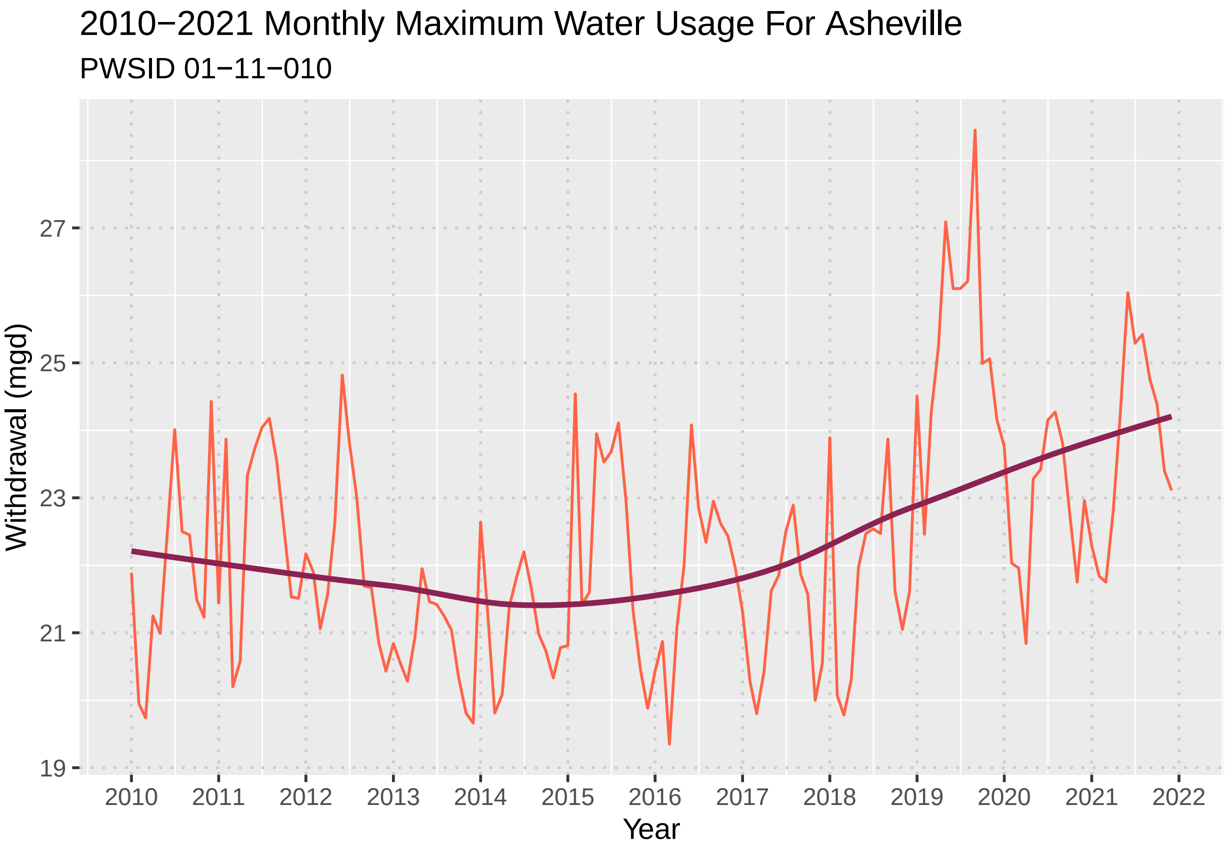


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2021.Add a smoothed line to the plot (method = 'loess').

   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
#subset the years
the_years = rep(2010:2021)
#"map" the scraping_function to retrieve data for all years
Asheville_2010_2021 <- map2(the_years, '01-11-010', scraping_function)
#conflate the returned list of dataframes into a single one
df_Asheville_2010_2021 <- bind_rows(Asheville_2010_2021)

#plot the monthly data from 2010 to 2021
ggplot(df_Asheville_2010_2021,
       aes(x = Date, y = Maximum_Day)) +
  geom_line(color = 'tomato1') +
  geom_smooth(method = 'loess', se = FALSE, color = 'violetred4') +
  labs(title = "2010-2021 Monthly Maximum Water Usage For Asheville",
       subtitle = "PWSID 01-11-010",
       y = "Withdrawal (mgd)",
       x = "Year") +
  scale_x_date(date_labels = "%Y", date_breaks = "1 year") +
  theme(panel.grid.major = element_line(color = "grey80", linetype = "dotted"))
```

```
## ‘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: > According to the plot, Asheville water usage has increased gradually over time. This trend is consistent with the assumption that the population has steadily increased from 2010 until 2021. Each year shows a mid-year peak, representing that water withdrawal is regularly higher in the summer than in the winter.