Assignment 10: Data Scraping

Siyu Dong

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(ggplot2)

#getwd()

mytheme <- theme_classic() +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
theme_set(mytheme)
```

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
Durham_LWSP_2022web <-
  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
Durham_LWSP_WaterSystem <- Durham_LWSP_2022web %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()
Durham_LWSP_WaterSystem
```

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

```
Durham_LWSP_PWSID <- Durham_LWSP_2022web %>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
  html_text()
Durham_LWSP_PWSID
```

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

```
Durham_LWSP_Ownership <- Durham_LWSP_2022web %>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%
  html_text()
Durham_LWSP_Ownership
```

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

```
Durham_LWSP_MaxWithdraws <- Durham_LWSP_2022web %>%
  html_nodes("th~ td+ td") %>%
  html_text()
Durham_LWSP_MaxWithdraws
```

```
##  [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"
```

```
class(Durham_LWSP_MaxWithdraws)
```

```
## [1] "character"
```

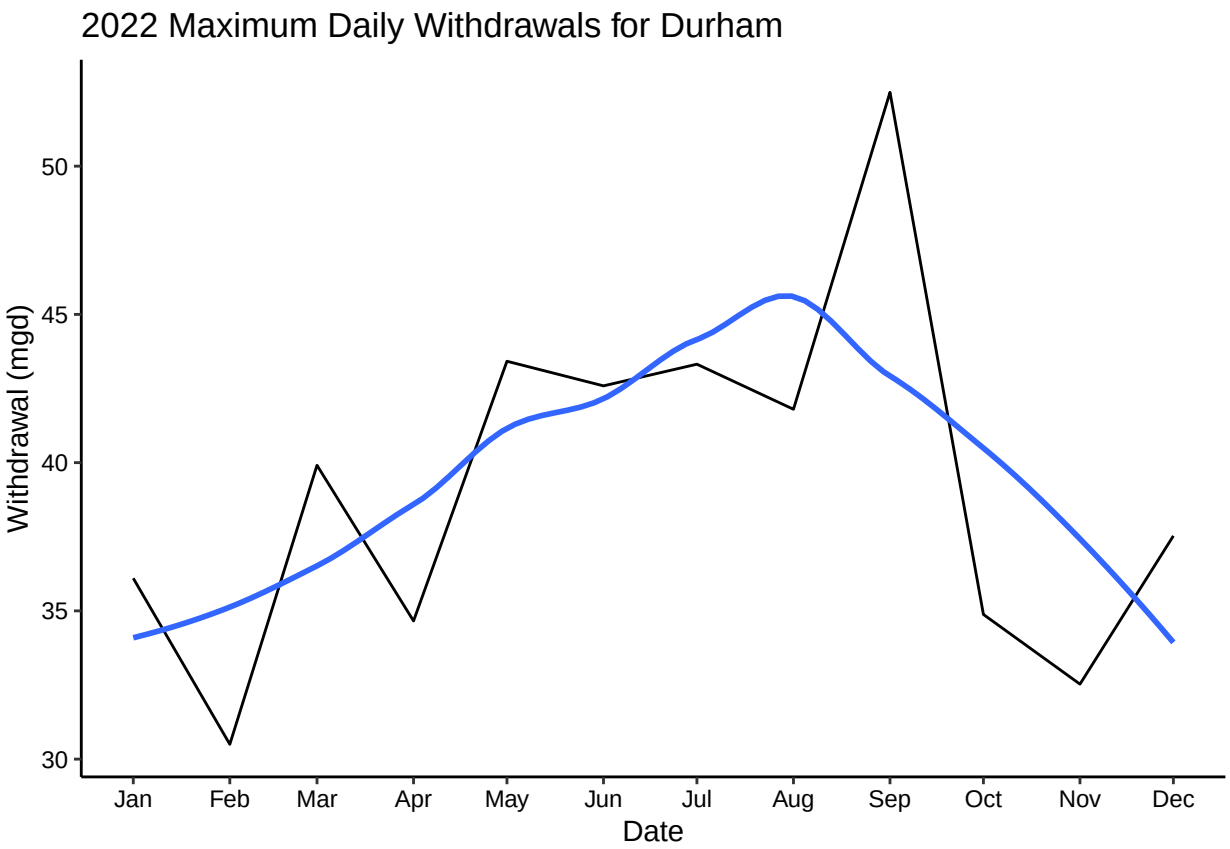
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   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
#Since withdraws were not scraped in chronological order,
#the corresponding month vector will be scraped here.
LWSP_Month <- Durham_LWSP_2022web %>%
  html_nodes(".fancy-table:nth-child(31) tr+ tr th") %>%
  html_text()

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

LWSP_Month_numeric <- as.numeric(match(LWSP_Month, months))

#Create a dataframe
Durham_LWSP_df <- data.frame("Year" = rep(2022, 12))

#Modify the dataframe
Durham_LWSP_df <- Durham_LWSP_df %>%
  mutate(WaterSystem_Name = !!Durham_LWSP_WaterSystem,
         PWSID = !!Durham_LWSP_PWSID,
         Ownership = !!Durham_LWSP_Ownership,
         MaxDayUse = as.numeric(Durham_LWSP_MaxWithdraws),
         Date = my(paste(LWSP_Month_numeric,"-",Year)))

#5
ggplot(Durham_LWSP_df, aes(x = Date, y = MaxDayUse)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2022 Maximum Daily Withdrawals for Durham"),
       y="Withdrawal (mgd)",
       x="Date") +
  scale_x_date(date_labels = "%b", 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**.

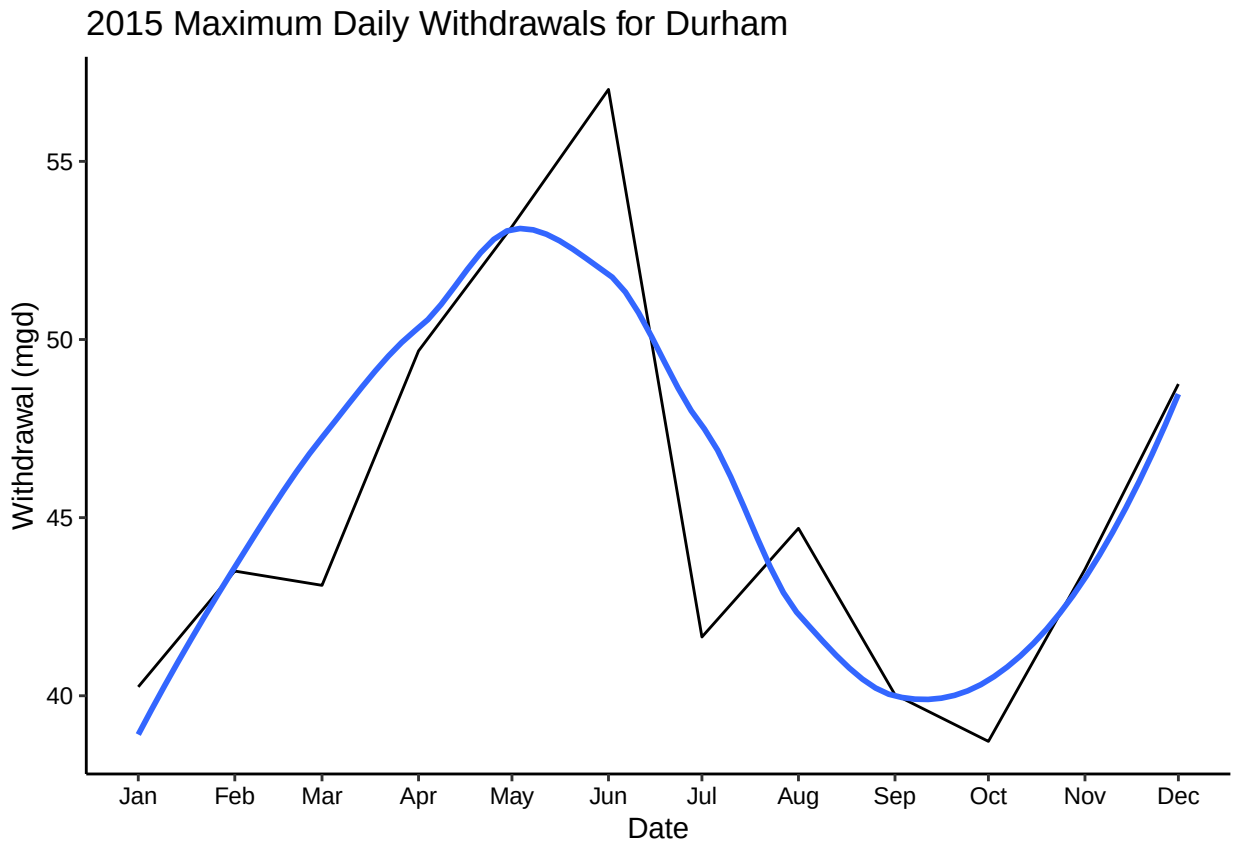
```
#6.
scrape.it <- function(the_year, the_pwsid){
  the_base_url <- 'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid='
  the_scrape_url <- paste0(the_base_url, the_pwsid, '&year=', the_year)

  the_website <- read_html(the_scrape_url)

  the_WaterSystem_tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'
  the_PWSID_tag <- 'td tr:nth-child(1) td:nth-child(5)'
  the_Ownership_tag <- 'div+ table tr:nth-child(2) td:nth-child(4)'
  the_MaxWithdraws_tag <- 'th~ td+ td'

  the_WaterSystem <- the_website %>% html_nodes(the_WaterSystem_tag) %>% html_text()
  the_PWSID <- the_website %>% html_nodes(the_PWSID_tag) %>%  html_text()
  the_Ownership <- the_website %>% html_nodes(the_Ownership_tag) %>% html_text()
  the_MaxWithdraws <- the_website %>% html_nodes(the_MaxWithdraws_tag) %>% html_text()

  df_withdrawals <- data.frame("Year" = rep(the_year,12)) %>%
    mutate(WaterSystem = !!the_WaterSystem,
           PWSID = !!the_PWSID,
           Ownership = !!the_Ownership,
```

```
            MaxDayUse = as.numeric(the_MaxWithdraws),
            Date = my(paste(LWSP_Month_numeric,"-",Year)))

  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

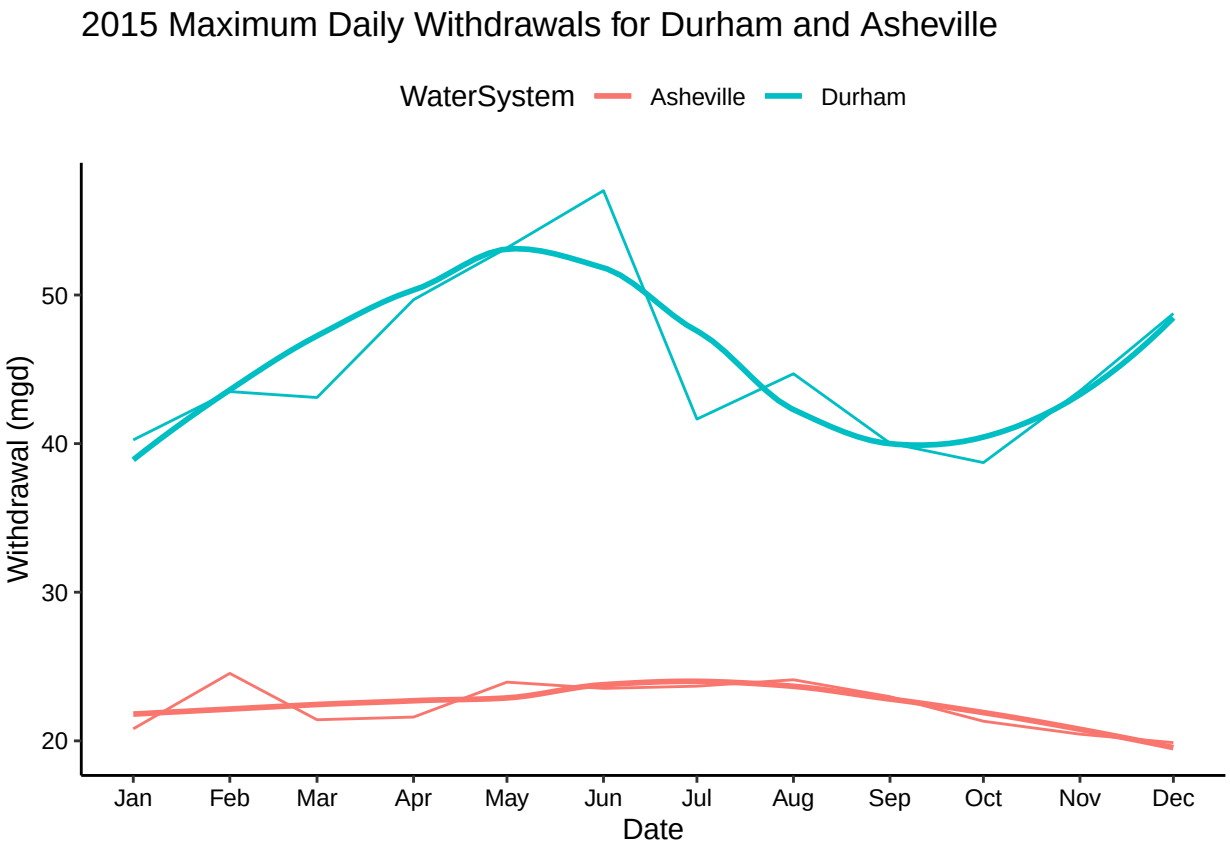
```
#7
Durham_2015 <- scrape.it(2015, '03-32-010')
view(Durham_2015)

ggplot(Durham_2015, aes(x = Date, y = MaxDayUse)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2015 Maximum Daily Withdrawals for Durham"),
       y="Withdrawal (mgd)",
       x="Date") +
  scale_x_date(date_labels = "%b", date_breaks = "1 month")
```

```
## ‘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.

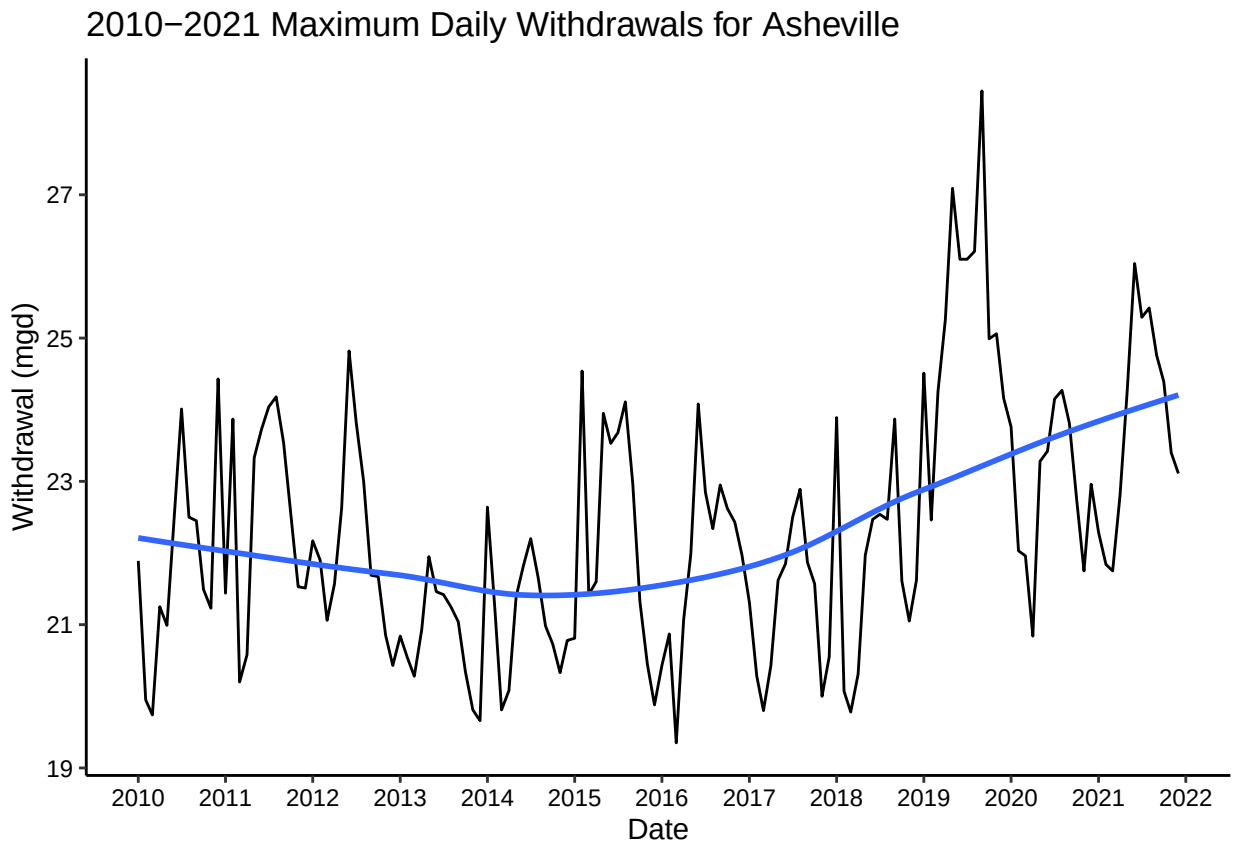
```
#8
Asheville_2015 <- scrape.it(2015, '01-11-010')
view(Asheville_2015)

Comparison_2015 <- rbind(Asheville_2015, Durham_2015)

ggplot(Comparison_2015, aes(x = Date, y = MaxDayUse, color = WaterSystem)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2015 Maximum Daily Withdrawals for Durham and Asheville"),
       y="Withdrawal (mgd)",
       x="Date") +
  scale_x_date(date_labels = "%b", 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.

```r
#9
the_years = rep(2010:2021)
my_pwsid = '01-11-010'

Asheville_dfs <- map2(the_years, my_pwsid, scrape.it)

Asheville_df <- bind_rows(Asheville_dfs)

ggplot(Asheville_df, aes(x = Date, y = MaxDayUse)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2010-2021 Maximum Daily Withdrawals for Asheville"),
       y="Withdrawal (mgd)",
       x="Date") +
  scale_x_date(date_labels = "%Y", date_breaks = "1 year")
```



Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time?

> Answer: With the year of 2015 as the turning point, before 2015, the water usage in Asheville slightly decreased; after 2015, it increased with a relatively rate.