Assignment 10: Data Scraping

Sophie Valkenberg

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
#Install familiar packages
library(tidyverse);library(lubridate);library(viridis);library(here)
here()

#install.packages("rvest")
library(rvest)

# Set theme
mytheme <- theme_gray() +
  theme(axis.text = element_text(angle = 45, color = "blue"),
        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 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 URL to be scraped. (In other words, read the contents into an `rvest` webpage object.)

```r
#2
theURL <- 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)“.

```r
#3
WaterSystemName<- 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()
Ownership <- theURL %>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>% html_text()
MaximumDayUse <- theURL %>%
  html_nodes("th~ td+ td") %>% html_text()

MaximumDayUse = as.numeric(MaximumDayUse)
```

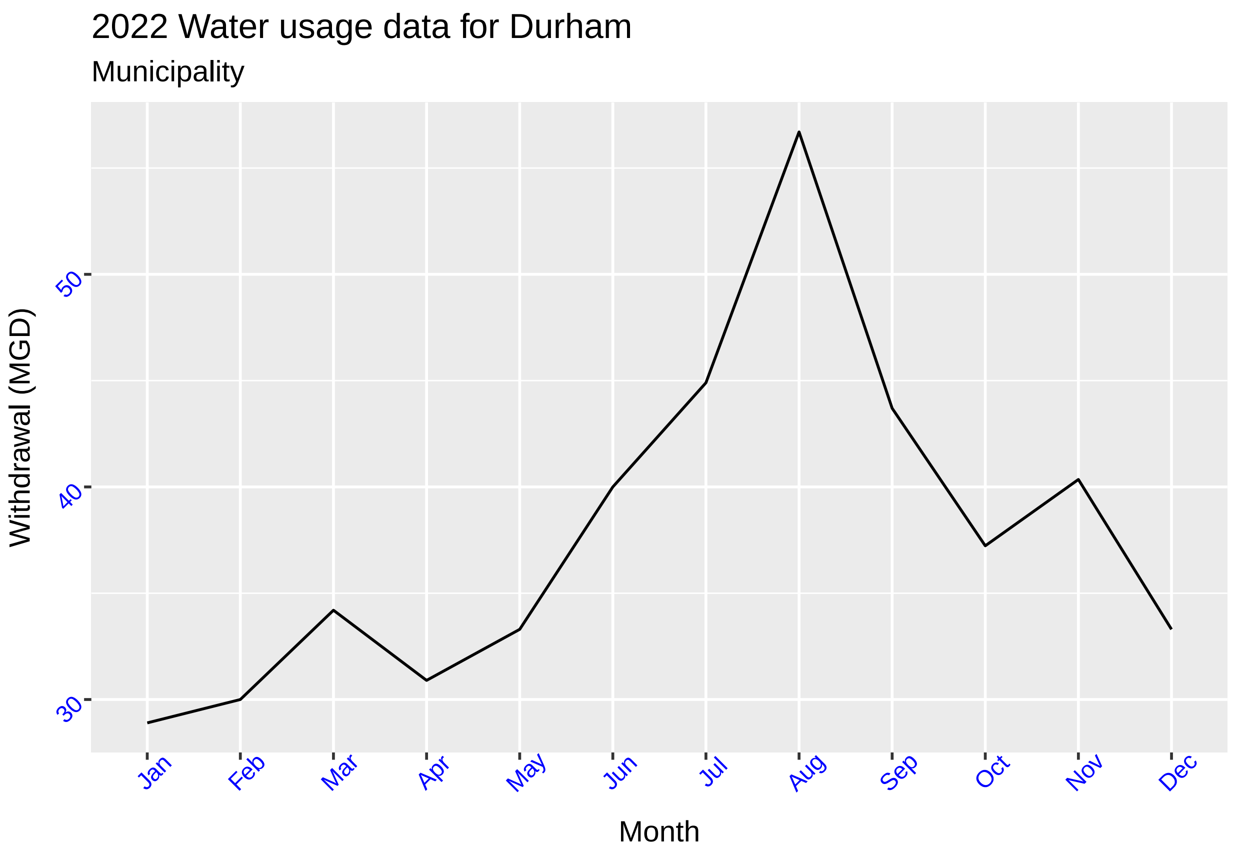
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in date format. (Feel free to add a Year column too, if you wish.)

   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
Months <- theURL %>%
  html_nodes(".fancy-table:nth-child(31) tr+ tr th") %>% html_text()

LWSP_dataframe <- data.frame(
  "WSName" = WaterSystemName,
  "PWSID" = PWSID,
  "Ownership" = Ownership,
  "Month" = as.factor(Months),
  "Maximim_Day_Use" = MaximumDayUse,
  "Year" = 2023
)

LWSP_dataframe <- LWSP_dataframe %>%
  mutate(Month = factor(Month, levels =
                          c("Jan","Feb","Mar","Apr","May","Jun","Jul","Aug",
                            "Sep","Oct","Nov","Dec"))) %>%
  arrange(Month) %>%
  mutate(Date = my(paste0(Month, "-", Year)))

#5
ggplot(LWSP_dataframe,aes(x=Month,y=Maximim_Day_Use, group =1)) +
  geom_line() +
  labs(title = paste("2022 Water usage data for", WaterSystemName),
       subtitle = Ownership,
       y="Withdrawal (MGD)",
       x="Month")
```



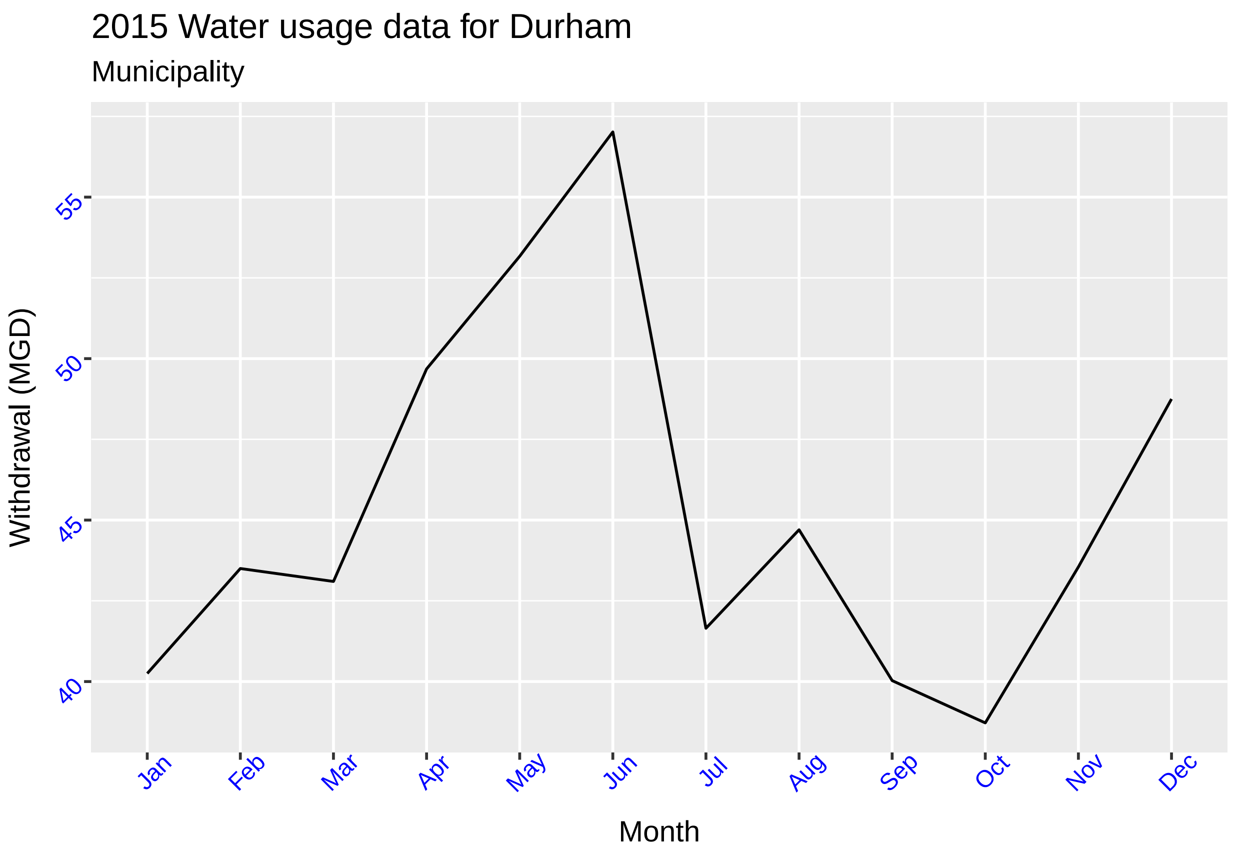
6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data, returning a dataframe. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

```
#6.
scrape.it <- function(the_PWSID, the_year){

  #Retrieve the website contents
  theURL <- read_html(paste0(
    'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',
    the_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)"
  MaximumDayUse_tag <- "th~ td+ td"

  #Scrape the data items
  WaterSystemName<- theURL %>%
  html_nodes(WaterSystemName_tag) %>% html_text()
  PWSID<- theURL %>%
  html_nodes(PWSID_tag) %>% html_text()
  Ownership <- theURL %>%
```

```r
  html_nodes(Ownership_tag) %>% html_text()
  MaximumDayUse <- theURL %>%
  html_nodes(MaximumDayUse_tag) %>% html_text()

  #Convert to a dataframe
 LWSP_dataframe <- data.frame(
  "WSName" = WaterSystemName,
  "PWSID" = PWSID,
  "Ownership" = Ownership,
  "Month" = as.factor(Months),
  "Maximim_Day_Use" = as.numeric(MaximumDayUse),
  "Year" = the_year
  )
 LWSP_dataframe <- LWSP_dataframe %>%
  mutate(Month = factor(Month, levels =
                          c("Jan","Feb","Mar","Apr","May","Jun","Jul","Aug",
                            "Sep","Oct","Nov","Dec"))) %>%
  arrange(Month) %>%
  mutate(Date = my(paste0(Month, "-", Year)))

 #scraping etiquette
  Sys.sleep(1)

  #Return the dataframe
  return(LWSP_dataframe)
}
```

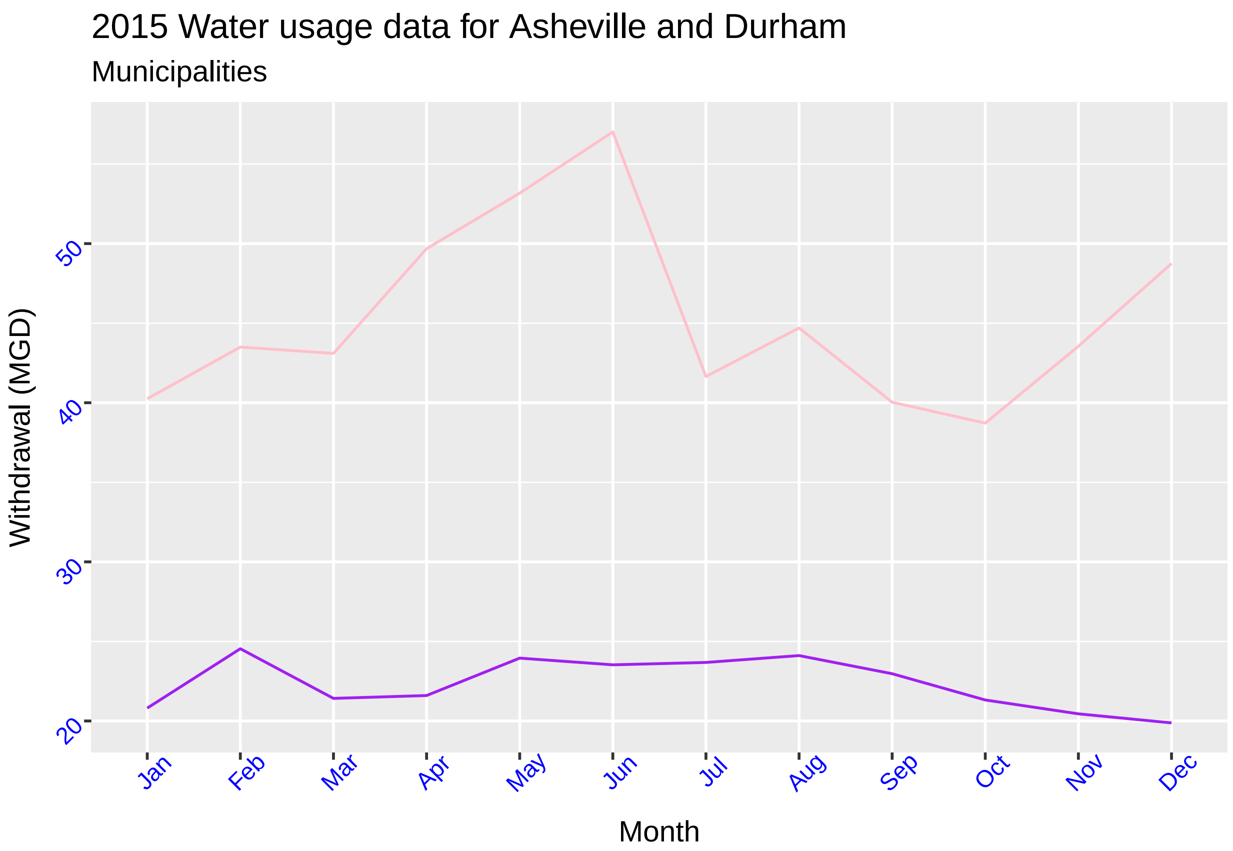
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID='03-32-010') for each month in 2015

```r
#7
df_2015_0332010 <- scrape.it("03-32-010", 2015)

ggplot(df_2015_0332010,aes(x=Month,y=Maximim_Day_Use, group = 1)) +
  geom_line() +
  labs(title = paste( df_2015_0332010$Year, "Water usage data for", df_2015_0332010$WSName),
       subtitle = df_2015_0332010$Ownership,
       y="Withdrawal (MGD)",
       x="Month")
```



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.

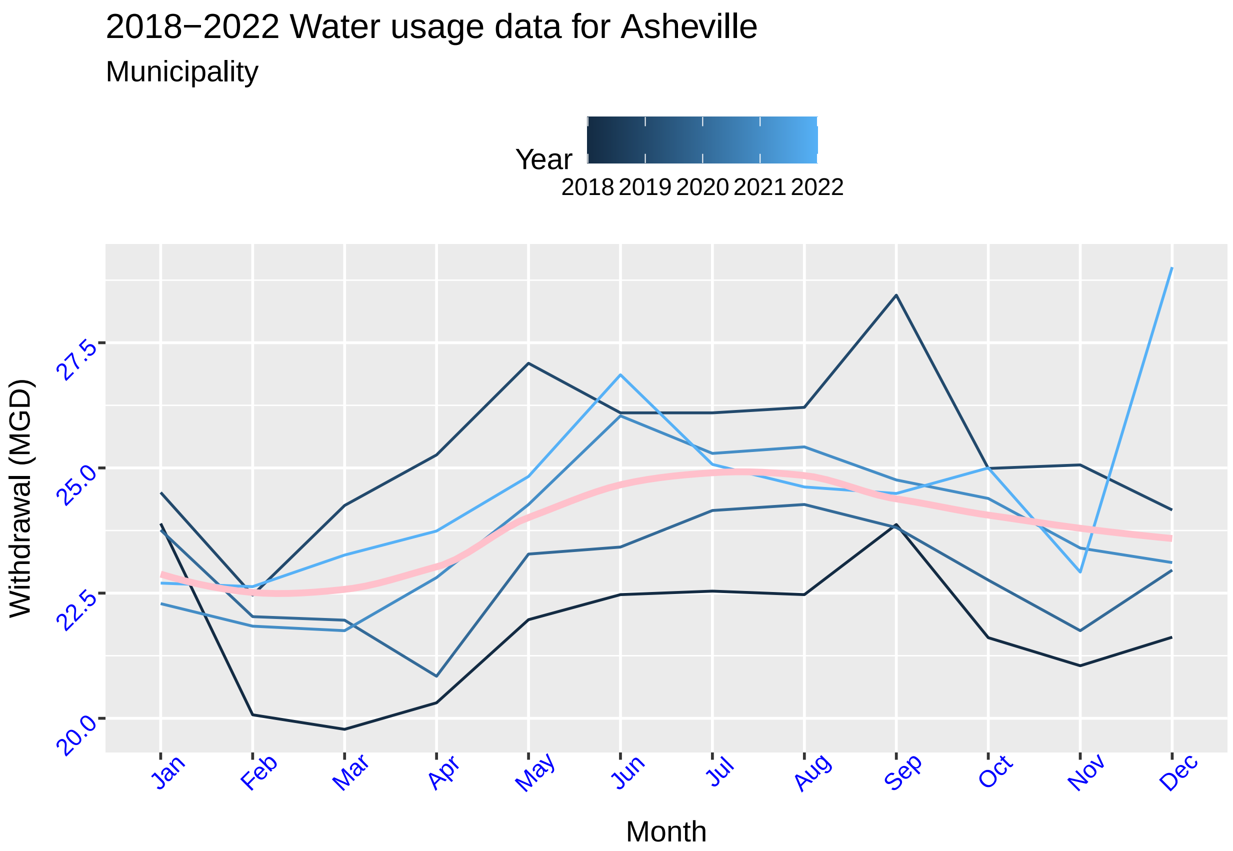
```
#8
df_2015_Asheville <- scrape.it("01-11-010", 2015)

ggplot() +
  geom_line(data = df_2015_0332010, aes(x=Month,y=Maximim_Day_Use, group = 1),
            color="pink") +
  geom_line(data = df_2015_Asheville, aes(x=Month,y=Maximim_Day_Use, group = 1),
            color="purple") +
  labs(title = paste("2015 Water usage data for Asheville and Durham"),
       subtitle = "Municipalities",
       y="Withdrawal (MGD)",
       x="Month")
```



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
the_facility_id <- "01-11-010"
the_years <- c(2018, 2019, 2020, 2021, 2022)
dfs_Asheville <- map2(the_facility_id, the_years, scrape.it)
df_Asheville <- bind_rows(dfs_Asheville)

ggplot(df_Asheville, aes(y = Maximim_Day_Use, x=Month, group=1)) +
  geom_line(aes(color = Year, group = Year) )+
  geom_smooth(method = "loess", se=FALSE, color = "pink", size = 1.2) +
  labs(title = paste("2018-2022 Water usage data for Asheville"),
       subtitle = "Municipality",
       y="Withdrawal (MGD)",
       x="Month")
```



Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time? > Answer: Asheville does have a trend in water usage over time. It seems that the municipality uses more water during the months of May-Oct/Nov than it does in Dec-Apr. This makes sense because of the temperature rise in the summer months and needing to water plants more frequently, hydrate more frequently, etc.