Assignment 09: Data Scraping

Stephanie Kinser

Total points:

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on data scraping.

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Fay_09_Data_Scraping.Rmd") prior to submission.

Set up

1. Set up your session:

- Check your working directory
- Load the packages `tidyverse`, `rvest`, and any others you end up using.
- Set your ggplot theme

```
#1

getwd()
```

```
## [1] "C:/Users/pogo/Documents/ENV872/Environmental_Data_Analytics_2022/Assignments"
```

```
library(tidyverse)
library(dplyr)
#install.packages("lubridate")
library(lubridate)
#install.packages("rvest")
library(rvest)

A9_theme <- theme_light(base_size = 12)+
   theme(axis.text = element_text(color = "black"),
        legend.position = "right", panel.grid.minor = element_blank())

theme_set(A9_theme)
```

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham's 2019 Municipal Local Water Supply Plan (LWSP):

- Navigate to https://www.ncwater.org/WUDC/app/LWSP/search.php
- Change the date from 2021 to 2020 in the upper right corner.

- 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=2020

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=2020')

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
- PSWID
- Ownership
- From the "3. Water Supply Sources" section:
- Average Daily Use (MGD) - for each month

In the code chunk below scrape these values, assigning them to three 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, with the first value being 36.0100.

```
#3
water.system.name <- webpage %>% html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>% html_text

pwsid <- webpage %>% html_nodes('td tr:nth-child(1) td:nth-child(5)') %>% html_text()

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

max.withdrawals.mgd <- webpage %>% html_nodes('th~ td+ td') %>% html_text()
```

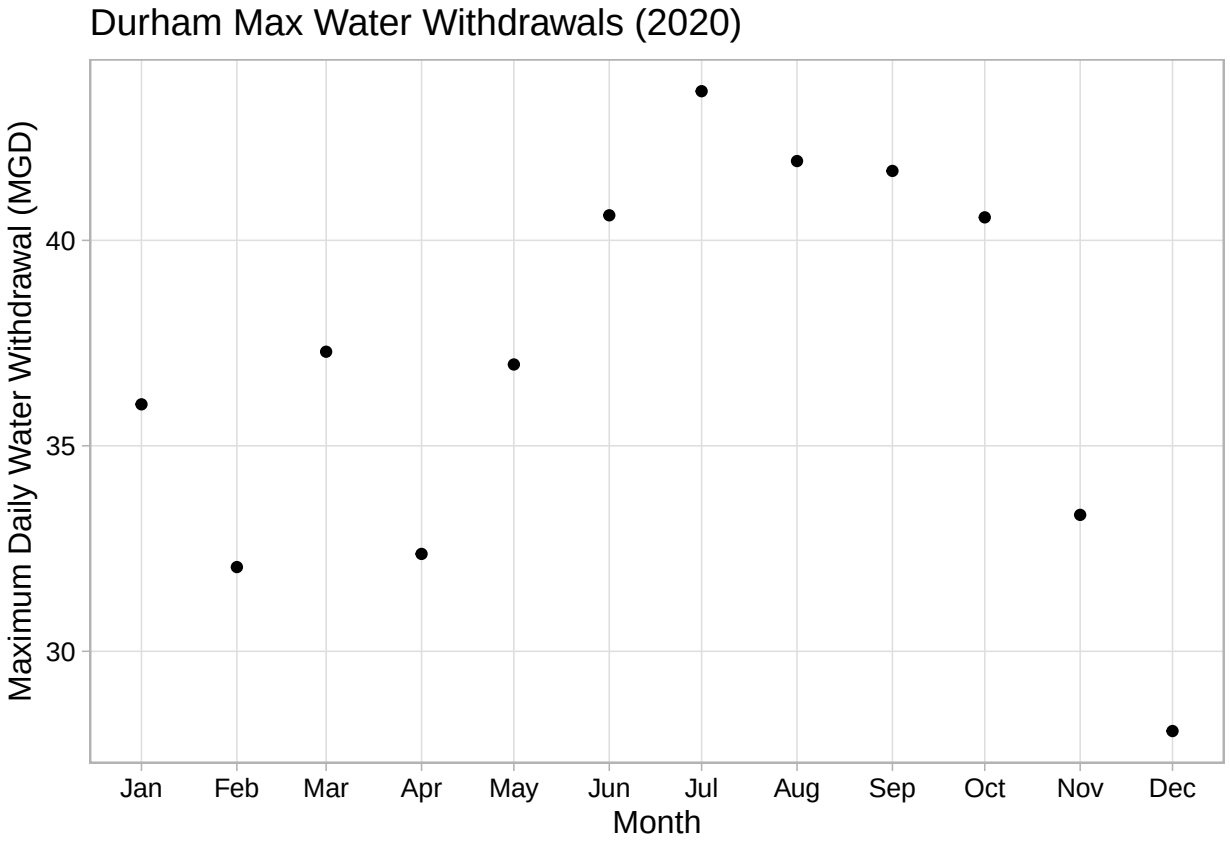
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in date format. (Feel free to add a Year column too, if you wish.)

   > 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 order. You can overcome this by creating a month column in the same order the data are scraped: Jan, May, Sept, Feb, etc...

5. Plot the max daily withdrawals across the months for 2020

```
#4
Month <- c(01, 05, 09, 02, 06, 10, 03, 07, 11, 04, 08, 12)
Year <- rep(2020, 12)

water_df <- data.frame(
```

```
  "Month" = Month,
  "Year" = Year,
  "Date" = my(paste(Month,"-",Year)),
  "Water_System" = rep(water.system.name, 12),
  "Owner" = rep(ownership, 12),
  "PWSID" = rep(pwsid, 12),
  "Max_Withdrawals" = round(as.numeric(max.withdrawals.mgd), 2)) %>%
  arrange(Date)


#5
ggplot(water_df, aes(x = Date, y = Max_Withdrawals)) +
  geom_point() +
  scale_x_date(date_labels = "%b", date_breaks = "1 month") +
  labs(title = "Durham Max Water Withdrawals (2020)", x = "Month", y = "Maximum Daily Water Withdrawal
```



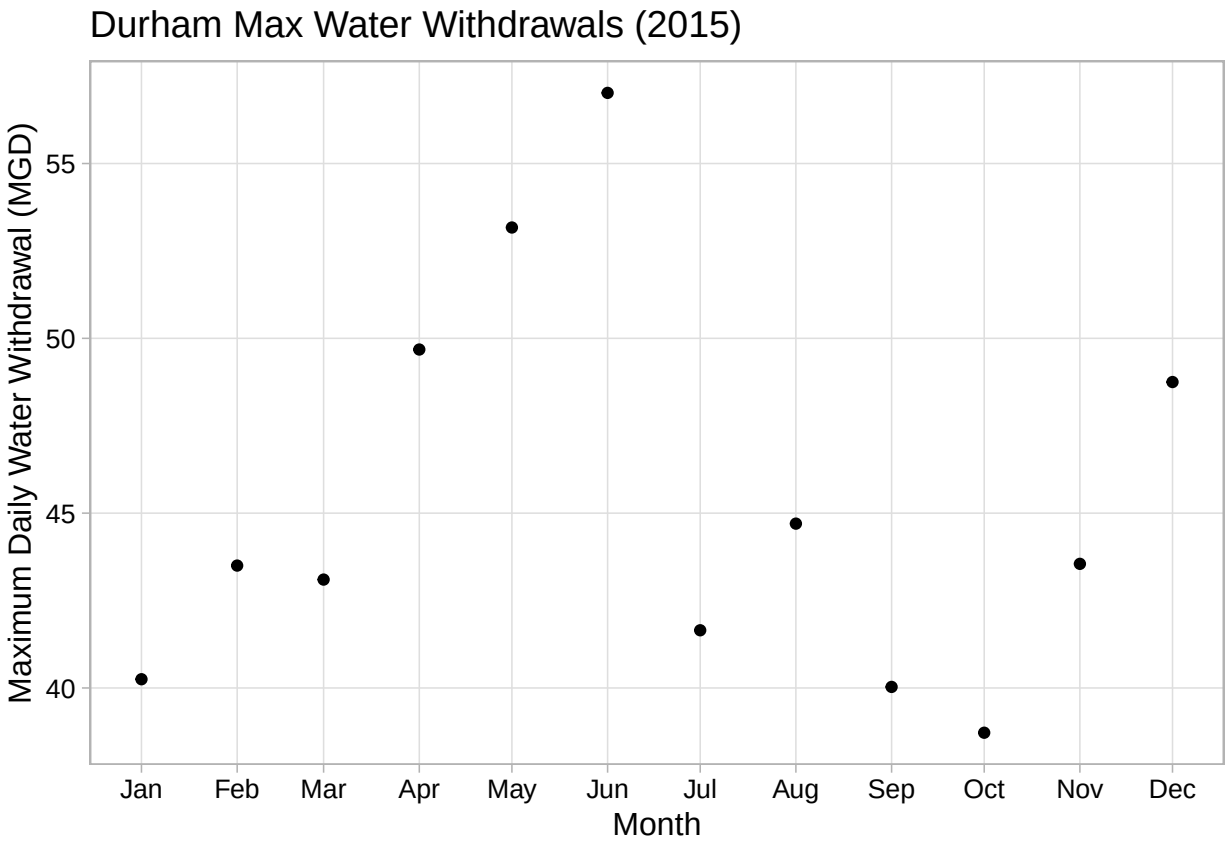
6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site scraped**.

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

  #Get the proper webpage
    webpage <- read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=', PWSID, '&yea
```

```
  #Locate elements and read their text attributes into variables
  water.system.name <- webpage %>% html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>% html_te

  pwsid <- webpage %>% html_nodes('td tr:nth-child(1) td:nth-child(5)') %>% html_text()

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

  max.withdrawals.mgd <- webpage %>% html_nodes('th~ td+ td') %>% html_text()


  #Construct a dataframe from the values
  Month <- c(01, 05, 09, 02, 06, 10, 03, 07, 11, 04, 08, 12)
  Year <- rep(the_year, 12)

  the_df <- data.frame(
    "Month" = Month,
    "Year" = Year,
    "Date" = my(paste(Month,"-",Year)),
    "Water_System" = rep(water.system.name, 12),
    "Owner" = rep(ownership, 12),
    "PWSID" = rep(pwsid, 12),
    "Max_Withdrawals" = round(as.numeric(max.withdrawals.mgd), 2)) %>%
    arrange(Date)

  return(the_df)
}
```

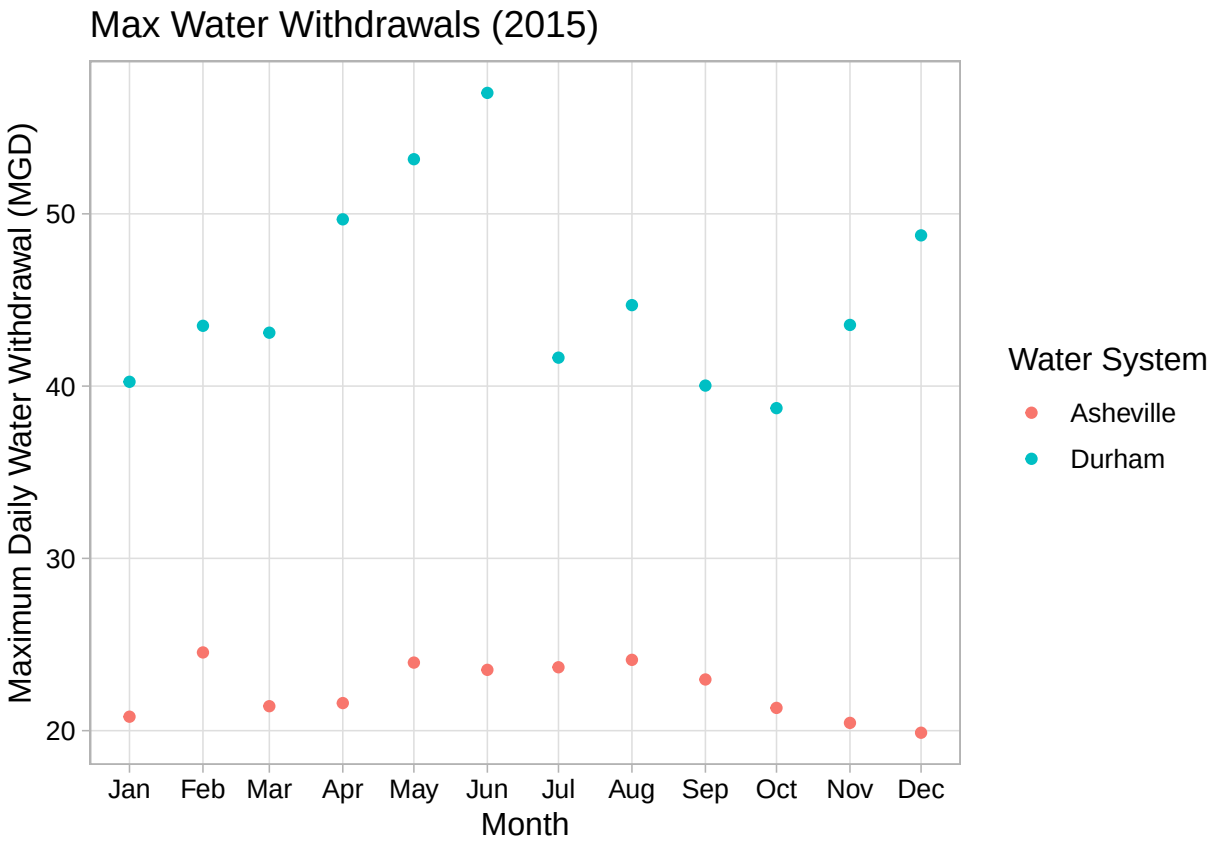
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID='03-32-010') for each month in 2015

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

ggplot(Durham_2015, aes(x = Date, y = Max_Withdrawals)) +
  geom_point() +
  scale_x_date(date_labels = "%b", date_breaks = "1 month") +
  labs(title = "Durham Max Water Withdrawals (2015)", x = "Month", y = "Maximum Daily Water Withdrawal
```



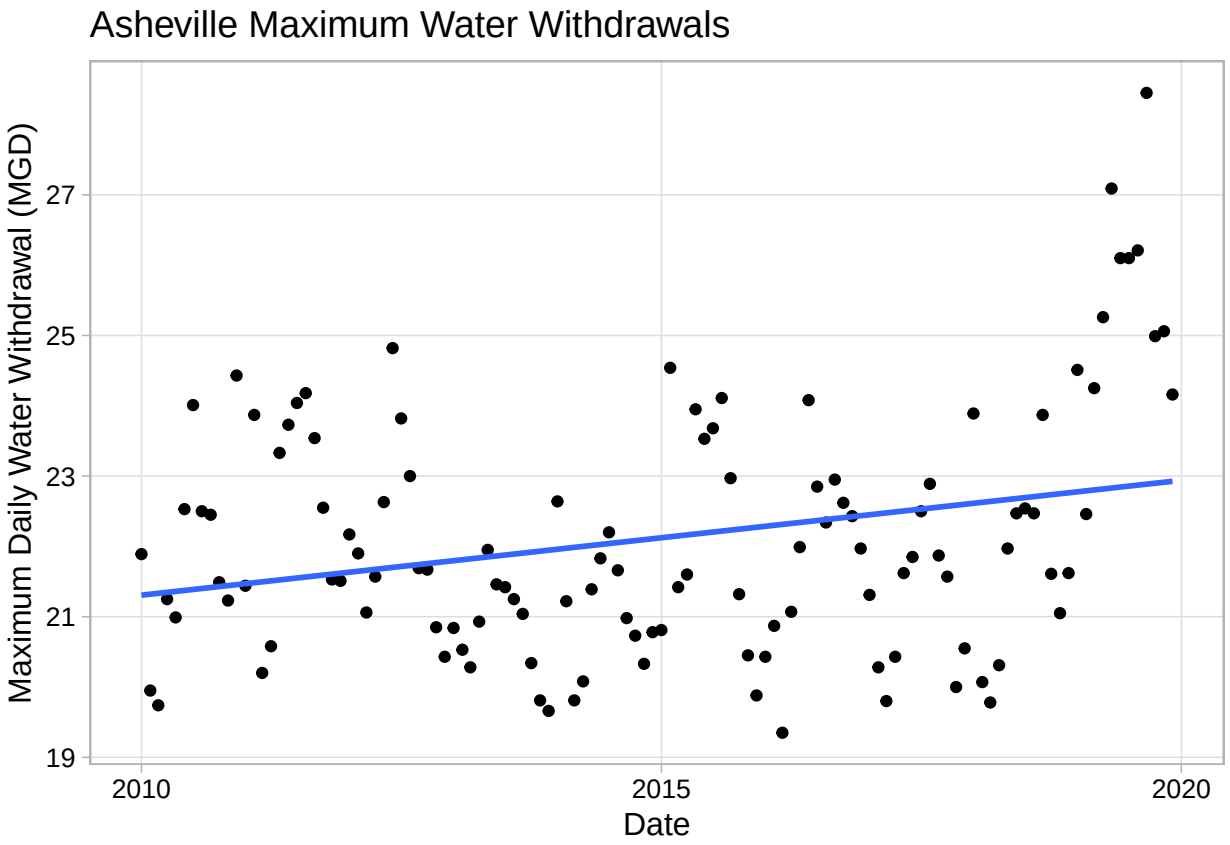
8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares the Asheville to Durham's water withdrawals.

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

water_2015 <- rbind(Durham_2015, Asheville_2015) %>%
  group_by(Water_System)

water_plot <- ggplot(water_2015)+
  geom_point(aes(x = Date, y = Max_Withdrawals, color = Water_System))+
  scale_x_date(date_labels = "%b", date_breaks = "1 month") +
  labs(title = "Max Water Withdrawals (2015)", x = "Month", y = "Maximum Daily Water Withdrawal (MGD)",

water_plot
```



9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

```
#9
Asheville_2010 <- scrape.it("01-11-010", 2010)
Asheville_2011 <- scrape.it("01-11-010", 2011)
Asheville_2012 <- scrape.it("01-11-010", 2012)
Asheville_2013 <- scrape.it("01-11-010", 2013)
Asheville_2014 <- scrape.it("01-11-010", 2014)
Asheville_2015 <- scrape.it("01-11-010", 2015)
Asheville_2016 <- scrape.it("01-11-010", 2016)
Asheville_2017 <- scrape.it("01-11-010", 2017)
Asheville_2018 <- scrape.it("01-11-010", 2018)
Asheville_2019 <- scrape.it("01-11-010", 2019)

Asheville_water <- rbind(Asheville_2010, Asheville_2011, Asheville_2012, Asheville_2013, Asheville_2014

Asheville_plot <- ggplot(Asheville_water, aes(x = Date, y = Max_Withdrawals)) +
  geom_point()+
  geom_smooth(method = lm, se = FALSE)+
  labs(title = "Asheville Maximum Water Withdrawals", x = "Date", y = "Maximum Daily Water Withdrawal (

Asheville_plot
```



Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time? The plot suggests that there has been an increase in maximum water usage from 2010 through 2019.