Assignment 09: Data Scraping

Addie Navarro

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] "Z:/EDA/Environmental_Data_Analytics_2022"
```

```
library(tidyverse)
library(rvest)
library(dplyr)
library(lubridate)
```

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 2020 to 2019 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
NC_Water_website <-
  read_html("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2020")
```

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:
- Max 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 <-
  NC_Water_website %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)")%>%
  html_text()

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

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

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

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

Year <- "2020"
```

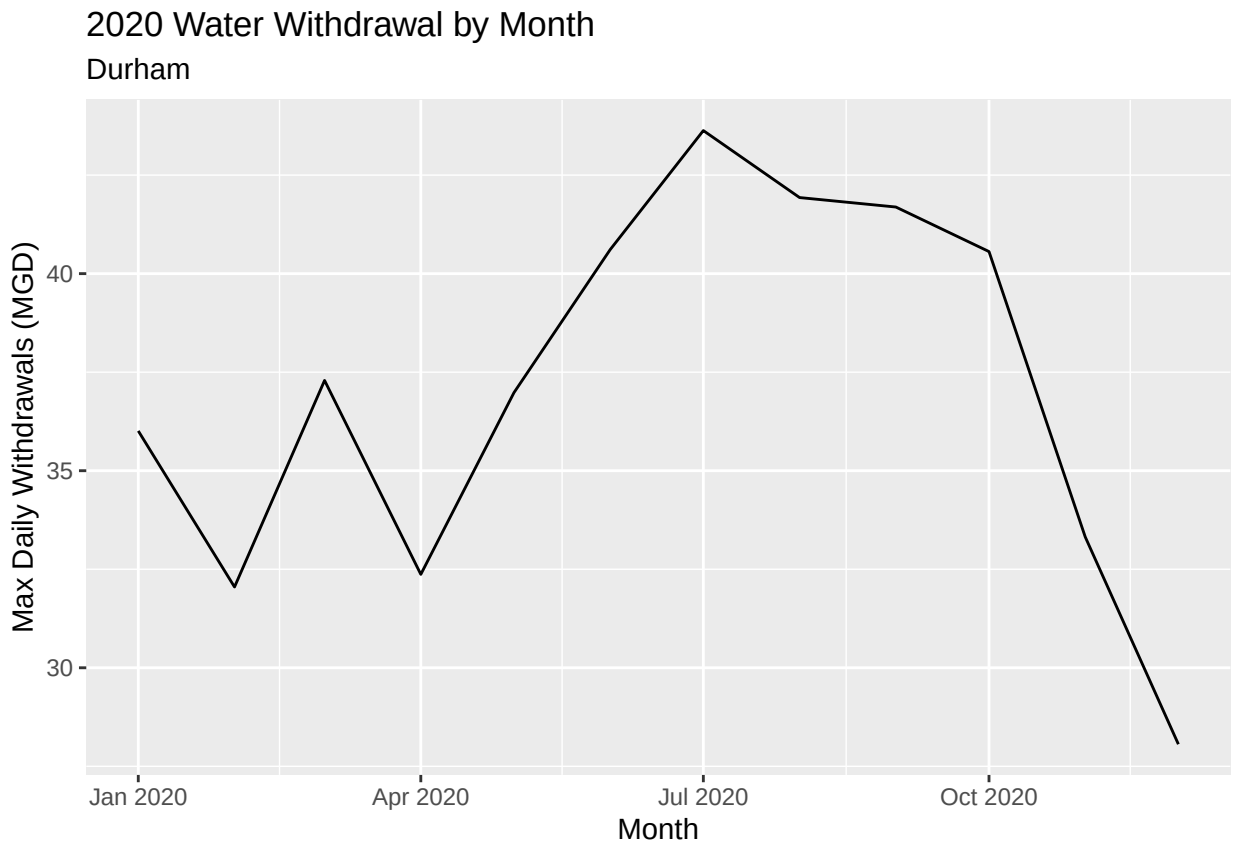
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   > 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 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

```r
#4
Water.df <-
  data.frame(
    "Water System" = water.system.name,
    "PWSID" = pwsid,
    "Ownership" = ownership,
    "Max Daily Use (MGD)" = as.numeric(max.withdrawals.mgd),
    "Month" = Month,
    "Year" = rep(Year, 12))%>%
  mutate(Date = my(paste(Month, "-", Year)))


#5
ggplot(Water.df,
       aes(x = Date, y = Max.Daily.Use..MGD.))+
  geom_line()+
  labs(title = paste(Year,"Water Withdrawal by Month"),
       subtitle = water.system.name,
       y = "Max Daily Withdrawals (MGD)",
       x = "Month")
```



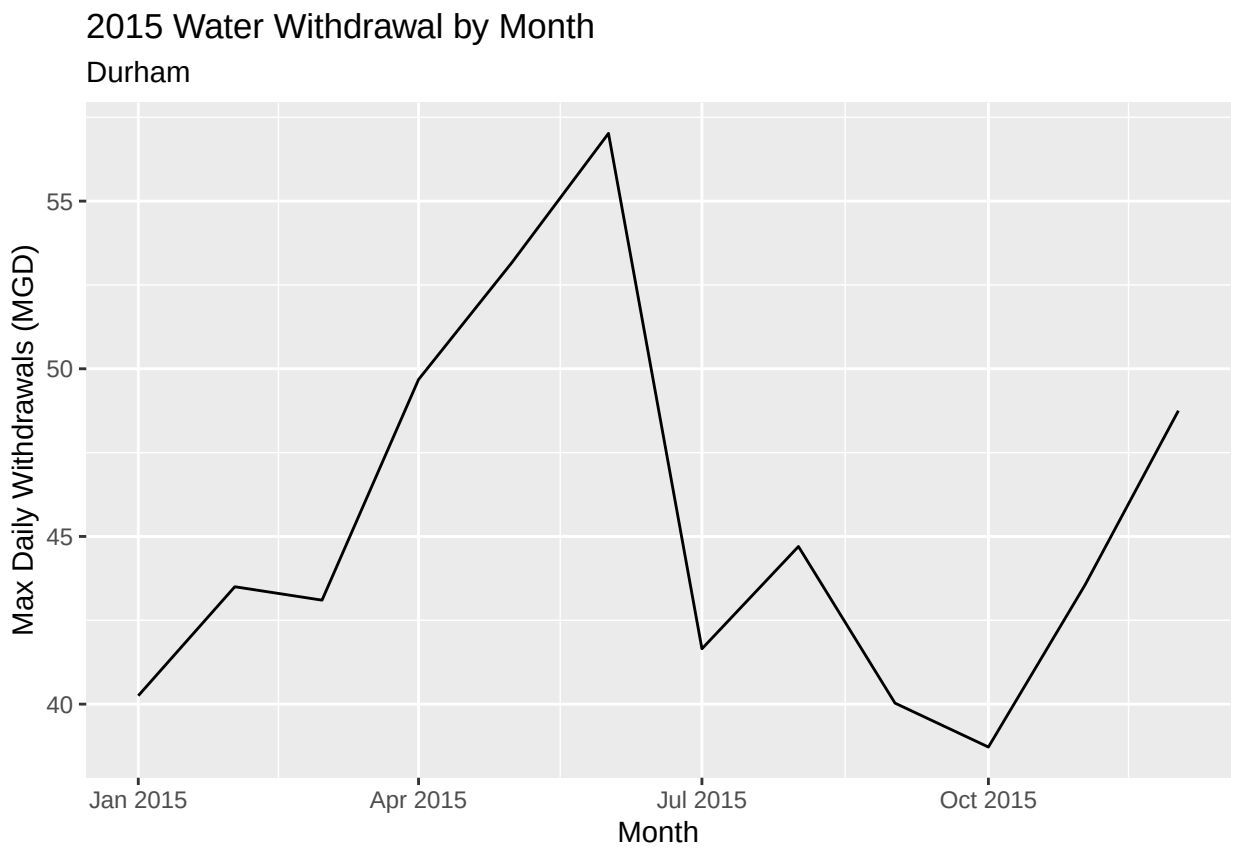
6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site scraped**.

```
#6.
NCwater_scrape <- function(the_year, pwsid){
  the_url <-
    read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',pwsid,'&year=',the_year,

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

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

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

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

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

Year <- the_year

the.df <-
  data.frame(
    "Water System" = rep(water.system.name, 12),
    "PWSID" = rep(pwsid, 12),
    "Ownership" = rep(ownership, 12),
    "Max Daily Use (MGD)" = as.numeric(max.withdrawals.mgd),
    "Month" = Month,
    "Year" = rep(Year, 12))%>%
  mutate(Date = my(paste(Month, "-", Year)))

 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

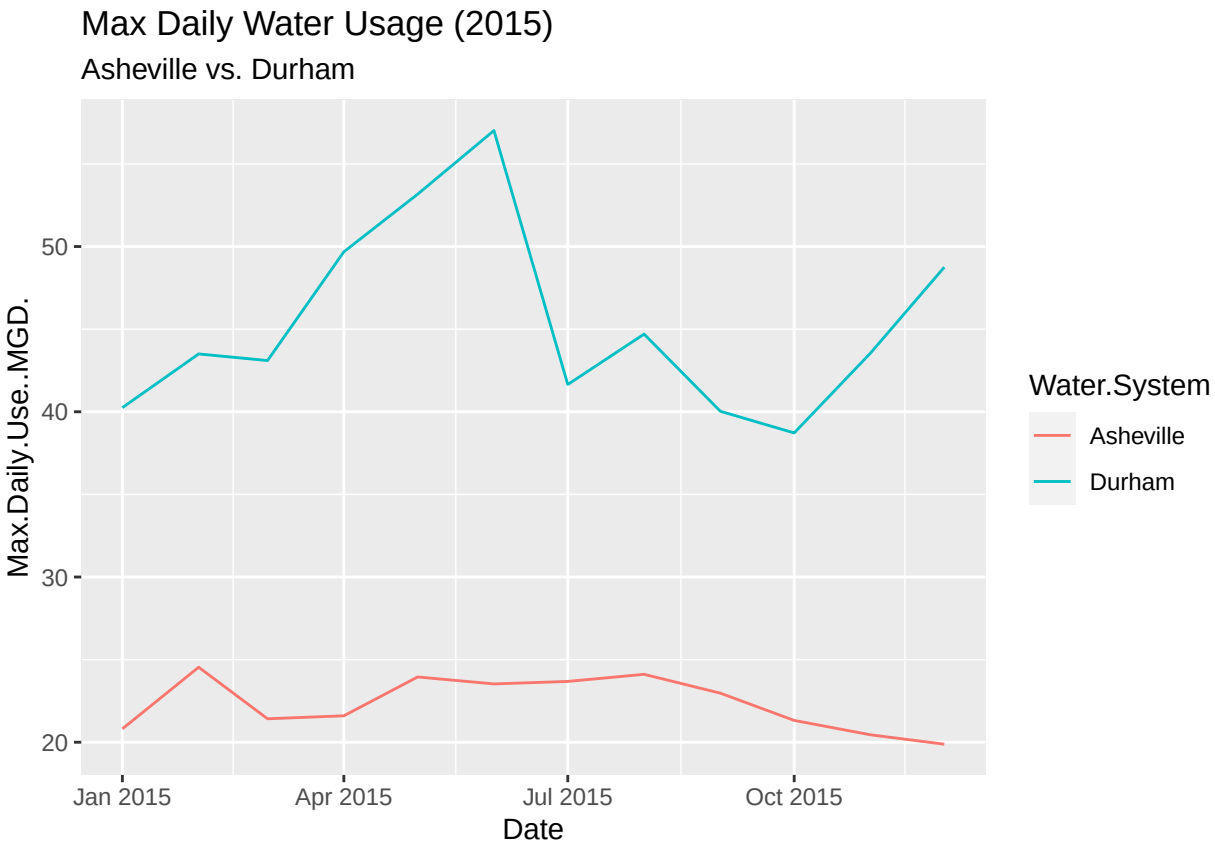
```
#7
Durham_2015 <- NCwater_scrape(2015, '03-32-010')

ggplot(Durham_2015,
       aes(x = Date, y = Max.Daily.Use..MGD.))+
  geom_line()+
    labs(title = "2015 Water Withdrawal by Month",
       subtitle = water.system.name,
       y = "Max Daily Withdrawals (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 the Asheville to Durham's water withdrawals.

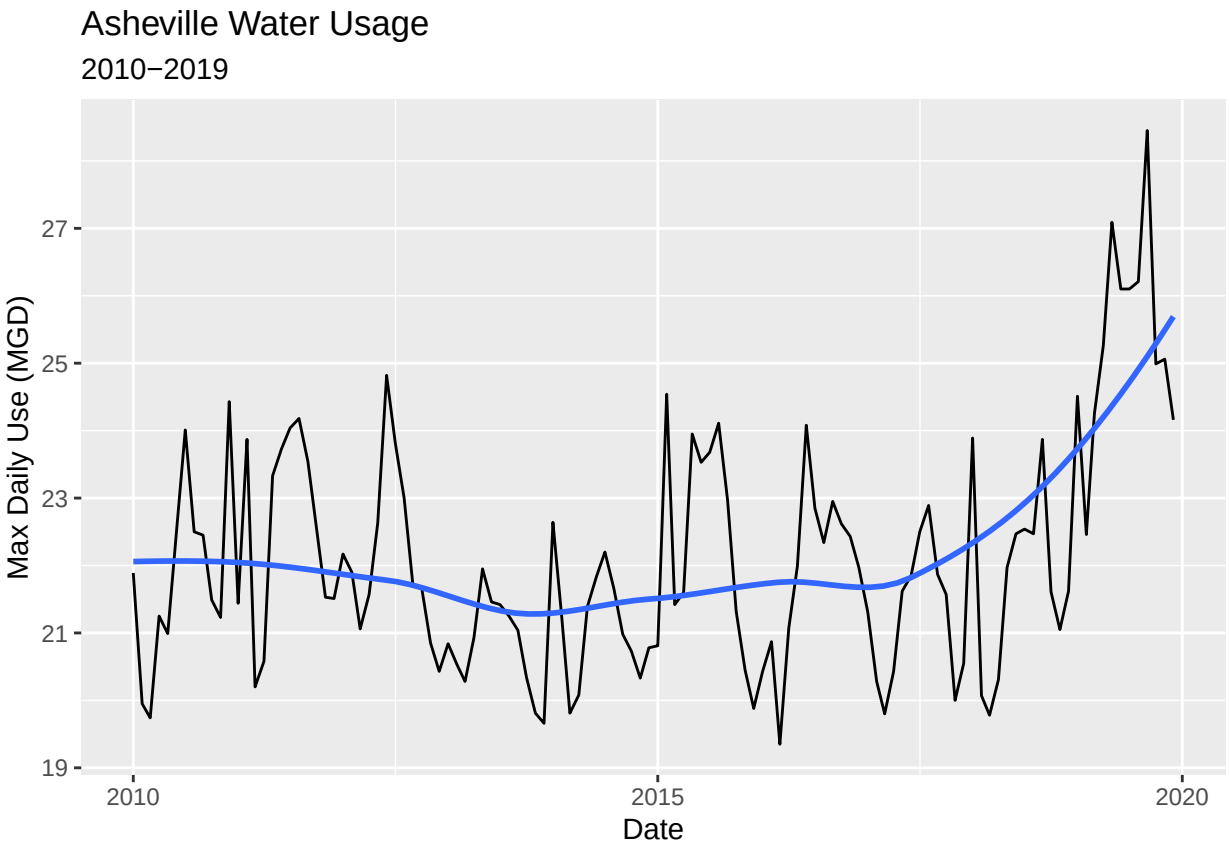
```
#8
Asheville_2015 <- NCwater_scrape(2015, '01-11-010')

#combine data with Durham 2015 data and plot
Ash_Durm_2015 <- bind_rows(Asheville_2015, Durham_2015)

ggplot(Ash_Durm_2015,
       aes(x = Date, y = Max.Daily.Use..MGD., color = Water.System))+
  geom_line()+
  labs(title = "Max Daily Water Usage (2015)",
       subtitle = "Asheville vs. Durham")
```



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
the_years = rep(2010:2019)
pwsid = '01-11-010'

Ash_decade <- lapply(X = the_years,
                     FUN = NCwater_scrape,
                     pwsid = pwsid)

A_decade.df <- bind_rows(Ash_decade)

ggplot(A_decade.df,
       aes(x = Date, y = Max.Daily.Use..MGD.))+
  geom_line()+
  geom_smooth(method="loess",se=FALSE)+
  labs(title = "Asheville Water Usage",
       subtitle = "2010-2019",
       x = "Date",
       y = "Max Daily Use (MGD)")
```

```
## `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?

Based on this plot, water usage has increased over time, but has gone up and down a lot over the years.