Assignment 09: Data Scraping

Jack Carpenter

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] "/Users/Jack/Documents/Duke/Spring 2022/Environmental Data Analytics/Environmental_Data_Analytics
```

```
library(tidyverse)
library(zoo)
library(lubridate)
library(viridis)
library(rvest)
library(dataRetrieval)


theme_set(theme_classic() +
  theme(axis.text = element_text(color = "black", size = 10),
        legend.position = "right"))
```

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 Read the website in
#theURL <- "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2020"

DurhamLWSP19_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:
- 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 Scrape it
water.system.name <- DurhamLWSP19_website %>%
  html_nodes("table:nth-child(7) tr:nth-child(1) td:nth-child(2)") %>%
  html_text()
pswid <- DurhamLWSP19_website %>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
  html_text()
ownership <- DurhamLWSP19_website %>%
  html_nodes("table:nth-child(7) tr:nth-child(2) td:nth-child(4)") %>%
  html_text()
max.withdrawals.mgd <- DurhamLWSP19_website %>%
  html_nodes(":nth-child(31) td:nth-child(9) , tr:nth-child(5) :nth-child(9) tr:nth-child(4) :nth-child
  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 data format. (Feel free to add a Year column too, if you wish.)

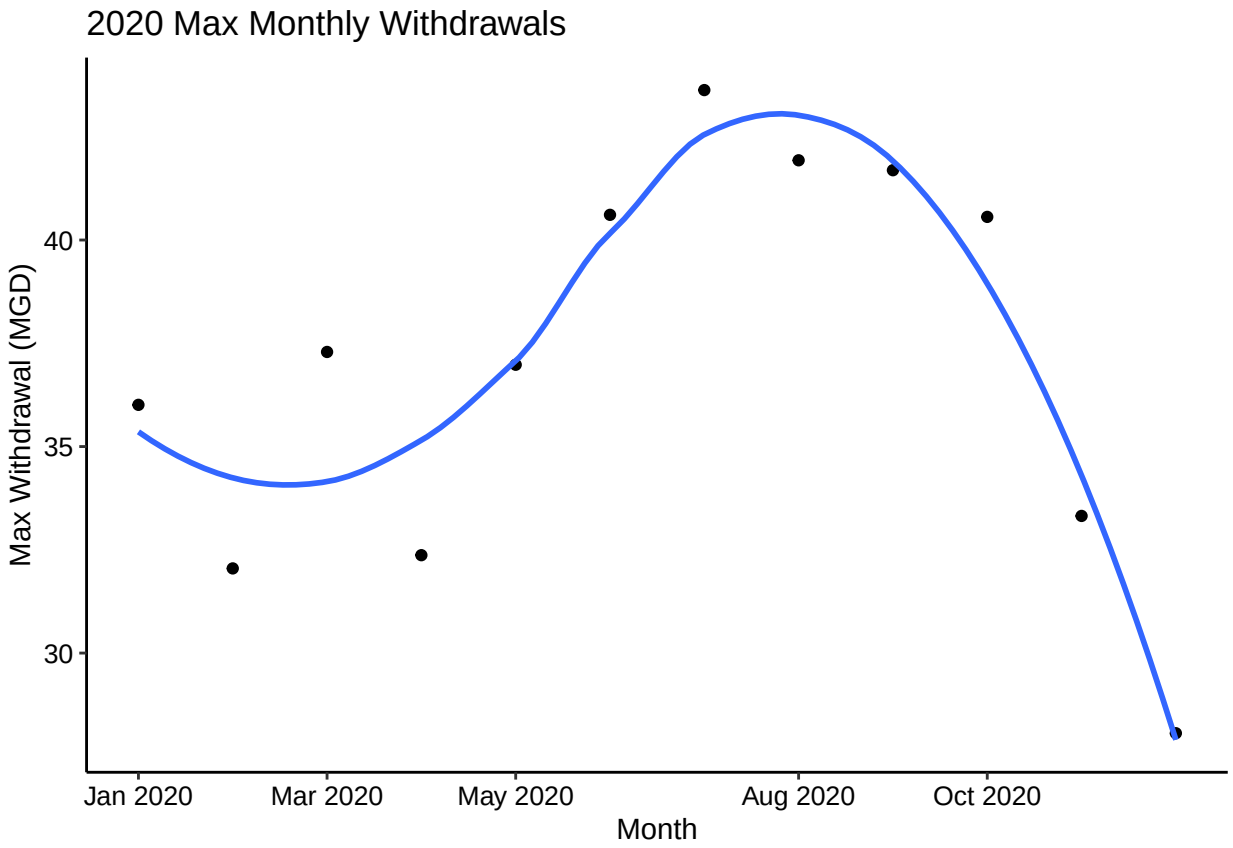
   > TIP: Use `rep()` to repeat a value when creating a dataframe.
   >
   > NOTE: It's likely you won't be able to scrape the monthly widthrawal data in order. You can overcome this by creating a month column in the same order the data are scraped: Jan, May, Sept, Feb, etc...

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

```
#4 Create a df
durhamwithdrawals_data <- data.frame("System_name" = water.system.name,
                                     "Ownership" = ownership,
                                     "PWSID" = as.character(pswid),
```

```
                                    "Max_Day_Withdrawals" =
                                      as.numeric(max.withdrawals.mgd),
                                    "Month" = c("Jan", "May", "Sep", "Feb",
                                                "Jun", "Oct", "Mar", "Jul",
                                                "Nov", "Apr", "Aug", "Dec"),
                                    "month_num" = c(1,5,9,2,6,10,3,7,11,4,8,12),
                                    "Year" = rep(2020,12))

durhamwithdrawals_data$Date <- as.yearmon(
  paste(durhamwithdrawals_data$Year, durhamwithdrawals_data$month_num),
  "%Y %m")

#durhamwithdrawals_data <- durhamwithdrawals_data %>%
  #mutate(date = make_date("Year", "month_num")) OR
  #mutate(date = my(paste(Month, "-", Year)))

#5 Plot withdrawals
ggplot(durhamwithdrawals_data, aes(x = Date, y = Max_Day_Withdrawals)) +
  geom_point() +
  geom_smooth(method = "loess", se = FALSE) +
  labs(x = "Month", y = "Max Withdrawal (MGD)",
       title = "2020 Max Monthly Withdrawals")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



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

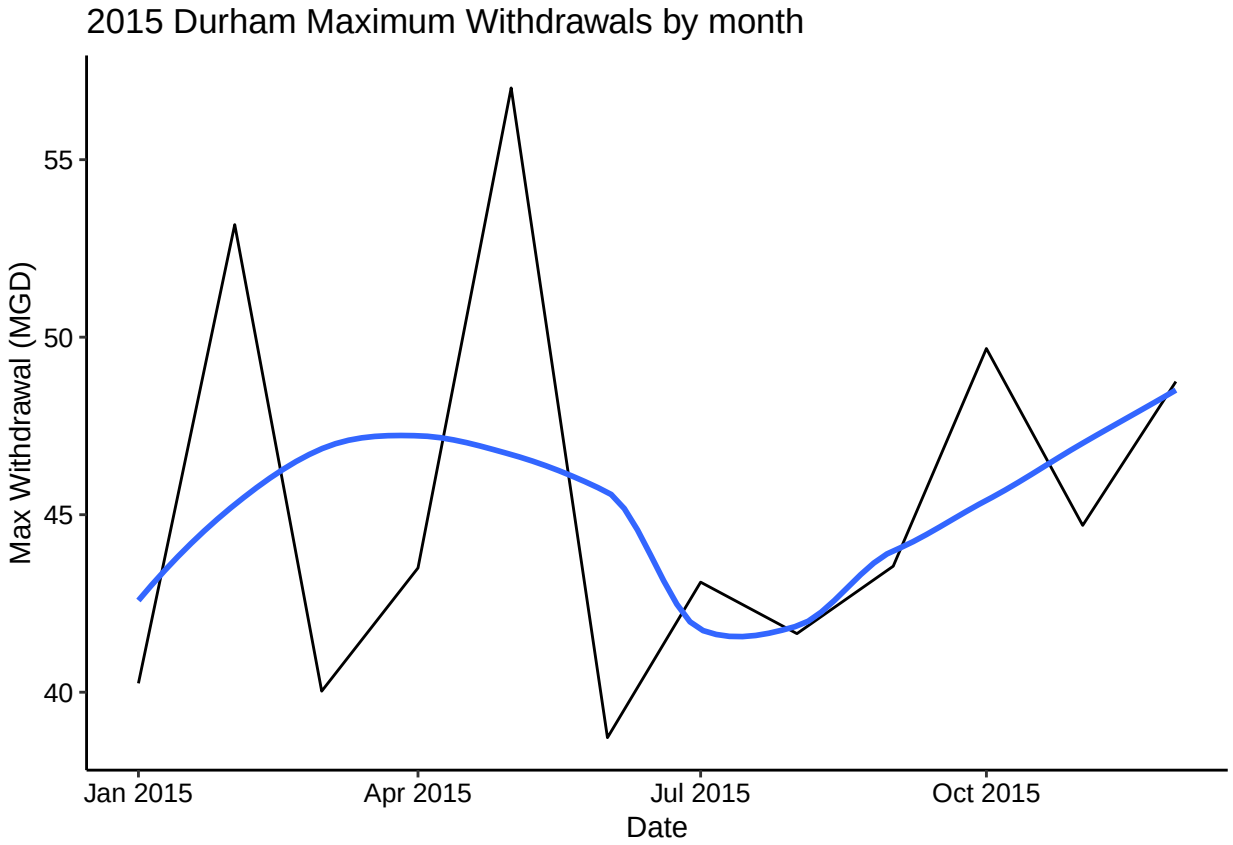
has data. **Be sure to modify the code to reflect the year and site scraped.**

```
#6. Make a scraping function
#first need to ID the needed pieces
the_year <- 2020
the_pwsid <- "03-32-010"
NCLWSP_scraper <- function(the_pwsid, the_year){

   #now link the website
  NCLWSP_website <- read_html(paste0(
    "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=",
    the_pwsid,"&year=",the_year))

   #set element variables
  system.name.tag <- "table:nth-child(7) tr:nth-child(1) td:nth-child(2)"
  pwsid.tag <- "td tr:nth-child(1) td:nth-child(5)"
  ownership.tag <- "table:nth-child(7) tr:nth-child(2) td:nth-child(4)"
  data.tag <- "th~ td+ td"

    #now name the elements
  water_system_name <- NCLWSP_website %>% html_nodes(system.name.tag) %>%
    html_text()
  system_pwsid <- NCLWSP_website %>% html_nodes(pwsid.tag) %>% html_text()
  system_ownership <- NCLWSP_website %>% html_nodes(ownership.tag) %>%
    html_text()
  max_withdrawals.mgd <- NCLWSP_website %>% html_nodes(data.tag) %>%
    html_text()

   #now make a dataframe
  NCLWSP_withdrawaldata <- data.frame("Max_withdrawals_mgd" =
                                        as.numeric(max_withdrawals.mgd),
                                      "Month" = rep(1:12),
                                     "Year" = rep(the_year,12)) %>%
    mutate(System_name = !!water_system_name,
           System_pwsid = !!system_pwsid,
           System_ownership = !!system_ownership,
           Date = my(paste(Month, "-", Year)))

  #Sys.sleep(5) #polite 5-second pause b/t requests, but not needed for this use

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

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

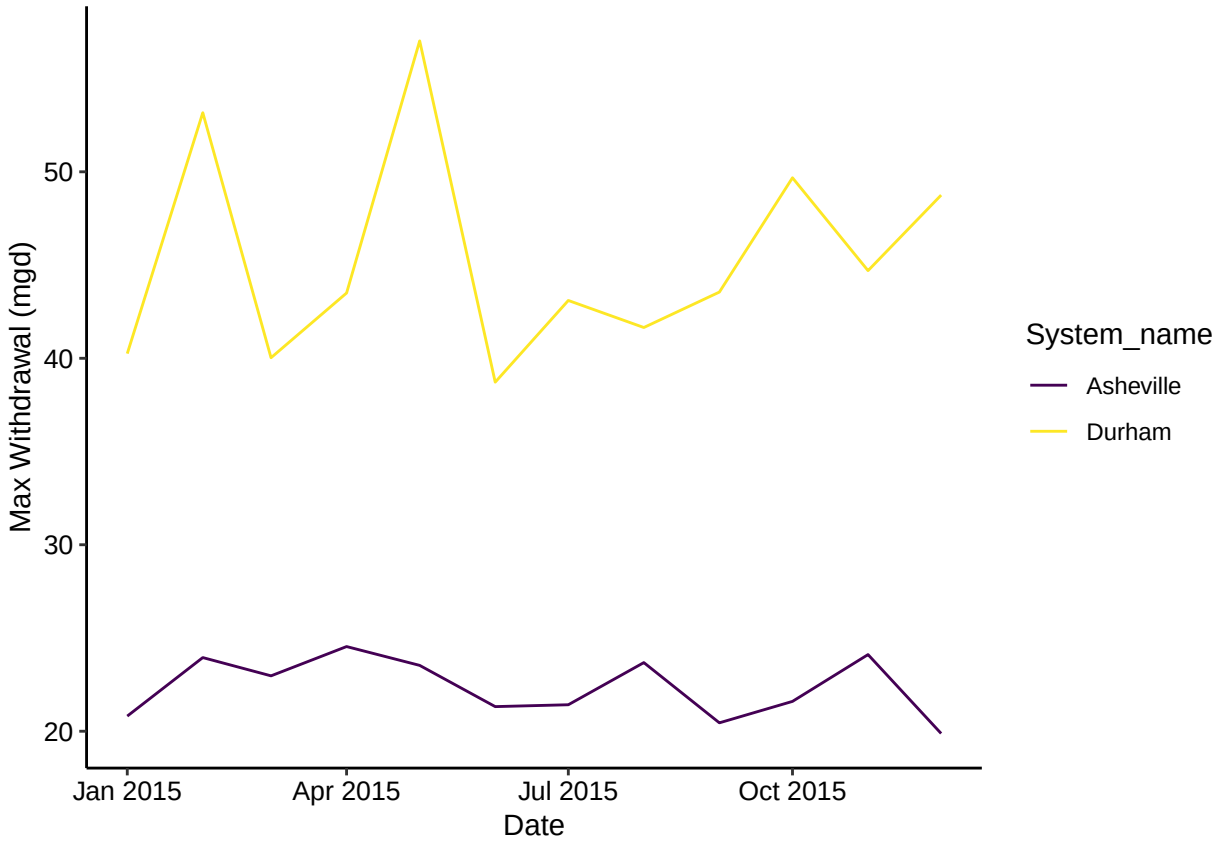
```
#7 Durham, 2015 data (PWSID = 03-32-010, year = 2015)
durham_2015 <- NCLWSP_scraper("03-32-010", 2015)
#view(durham_2015)

ggplot(durham_2015, aes(x = Date, y = Max_withdrawals_mgd)) +
  geom_line() +
  geom_smooth(method = "loess", se = FALSE) +
  labs(x = "Date", y = "Max Withdrawal (MGD)",
       title = "2015 Durham Maximum Withdrawals by 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 the Asheville to Durham's water withdrawals.

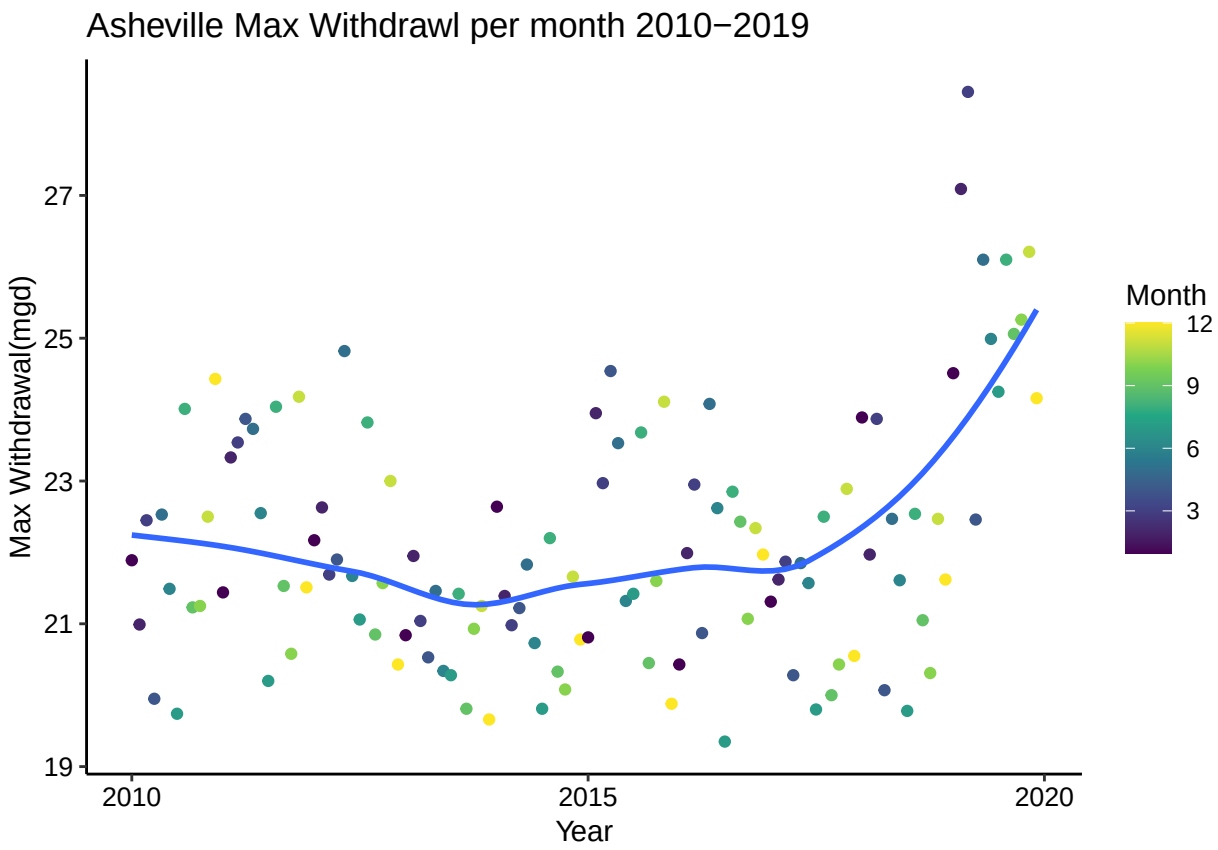
```
#8 Now to try it for Asheville
Asheville_2015 <- NCLWSP_scraper("01-11-010", 2015)
view(Asheville_2015)

Asheville_Durham_2015 <- bind_rows(durham_2015, Asheville_2015)

ggplot(Asheville_Durham_2015, aes(x = Date, y = Max_withdrawals_mgd,
                                  color = System_name)) +
  geom_line()+
  scale_color_viridis_d() +
  labs(x = "Date", y = "Max Withdrawal (mgd)")
```



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 through 2019
the_years = rep(2010:2019)
my_pwsid = "01-11-010"

Asheville_2010to2019 <- lapply(X = the_years,
                               FUN = NCLWSP_scraper,
                               the_pwsid = my_pwsid)

Asheville_bound <- bind_rows(Asheville_2010to2019)

ggplot(Asheville_bound, aes(x = Date, y = Max_withdrawals_mgd,
                            color = Month))+
  geom_point() +
  scale_color_viridis_c() +
  geom_smooth(method = "loess", se = FALSE) +
  labs(x = "Year", y = "Max Withdrawal(mgd)", color = "Month",
       title = "Asheville Max Withdrawl per month 2010-2019")
```

```
## `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? Yes, there appears to be an increasing trend in use over time.