Assignment 09: Data Scraping

Sam Vanasse

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/samvanasse/Desktop/ENV872-R/Environmental_Data_Analytics_2022/Assignments"
```

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

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

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
waterURL <-
  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
water.system.name <- waterURL %>%
  html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>% html_text()
pwsid <- waterURL %>%
  html_nodes('td tr:nth-child(1) td:nth-child(5)') %>% html_text()
ownership <- waterURL %>%
  html_nodes('div+ table tr:nth-child(2) td:nth-child(4)') %>% html_text()
max.withdrawals.mgd <- waterURL %>% html_nodes('th~ td+ td') %>% html_text()
```

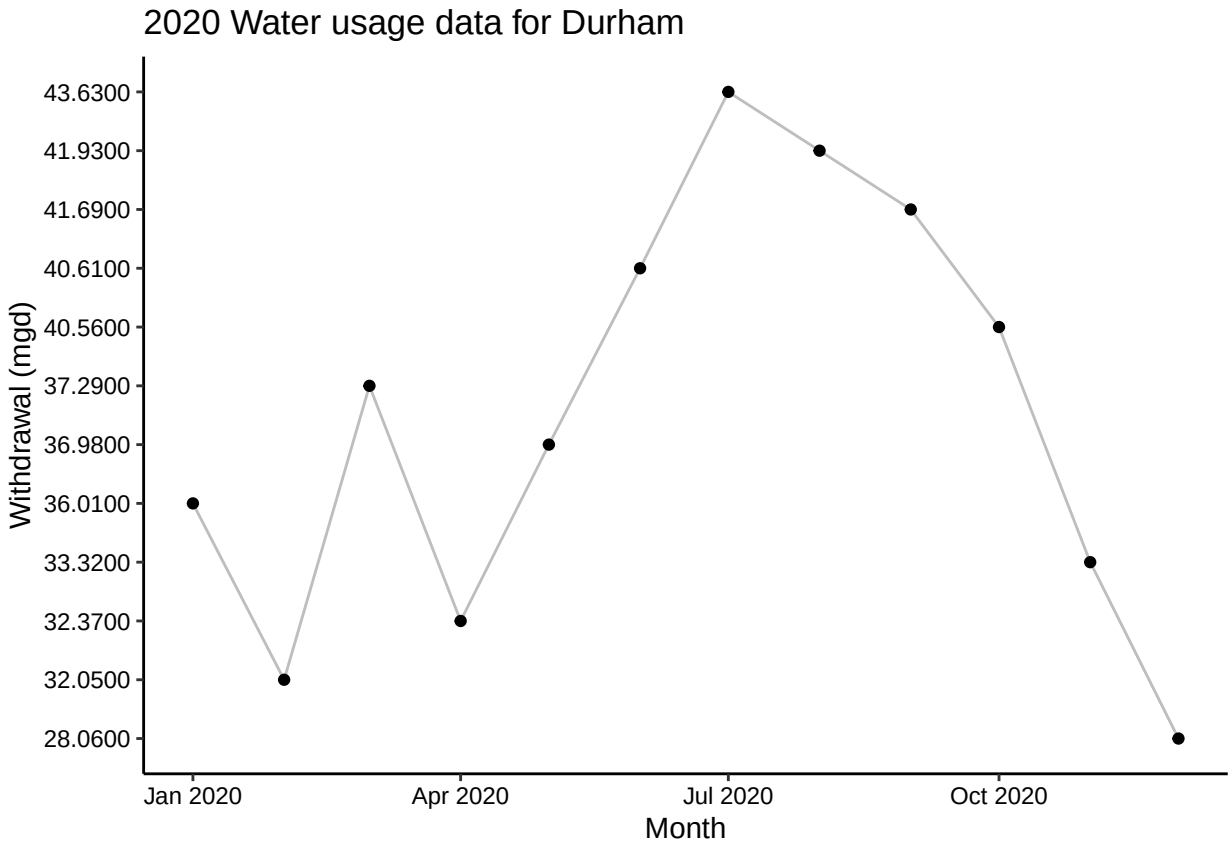
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   > 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
month <- c(1, 5, 9, 2, 6, 10, 3, 7, 11, 4, 8, 12)

water.df <- data.frame("Name" = water.system.name,
                       "PWSID" = pwsid,
                       "Ownership" = ownership,
                       "Withdrawals" = max.withdrawals.mgd,
                       "Month" = month,
                       "Year" = rep(2020, 12))

water.df <- water.df %>%
  mutate(Date = my(paste(Month,"-",Year)))

#5
```

```
ggplot(water.df,aes(x=Date,y=max.withdrawals.mgd, group=1)) +
  geom_line(color="grey") +
  geom_point() +
  labs(title = paste("2020 Water usage data for",water.system.name),
       y="Withdrawal (mgd)",
       x="Month")
```



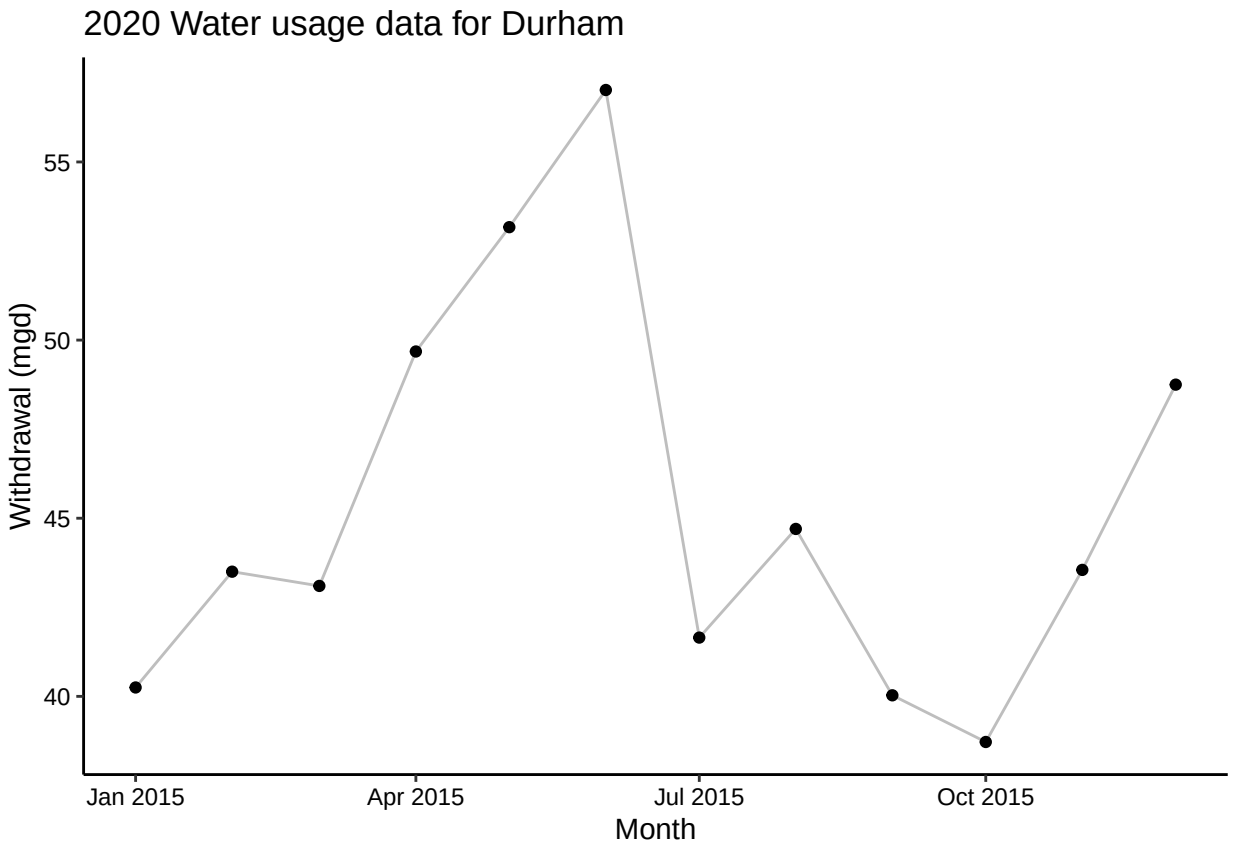
6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site scraped**.

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

   the_website <- read_html(
       paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',
              the_pwsid,'&year=',the_year))

  water.name.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)'
  withdrawals.tag <- 'th~ td+ td'

  water.system.name <- the_website %>%
    html_nodes(water.name.tag) %>% html_text()
  pwsid <- the_website %>%
    html_nodes(pwsid.tag) %>% html_text()
  ownership <- the_website %>%
```

```
    html_nodes(ownership.tag) %>% html_text()
  max.withdrawals.mgd <- the_website %>%
    html_nodes(withdrawals.tag) %>% html_text()
  month <- c(1, 5, 9, 2, 6, 10, 3, 7, 11, 4, 8, 12)

  the_df <-
    data.frame("Name" = as.character(water.system.name),
                       "PWSID" = as.character(pwsid),
                       "Ownership" = as.character(ownership),
                       "Withdrawals" = as.numeric(max.withdrawals.mgd),
                       "Month" = month,
                       "Year" = rep(the_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

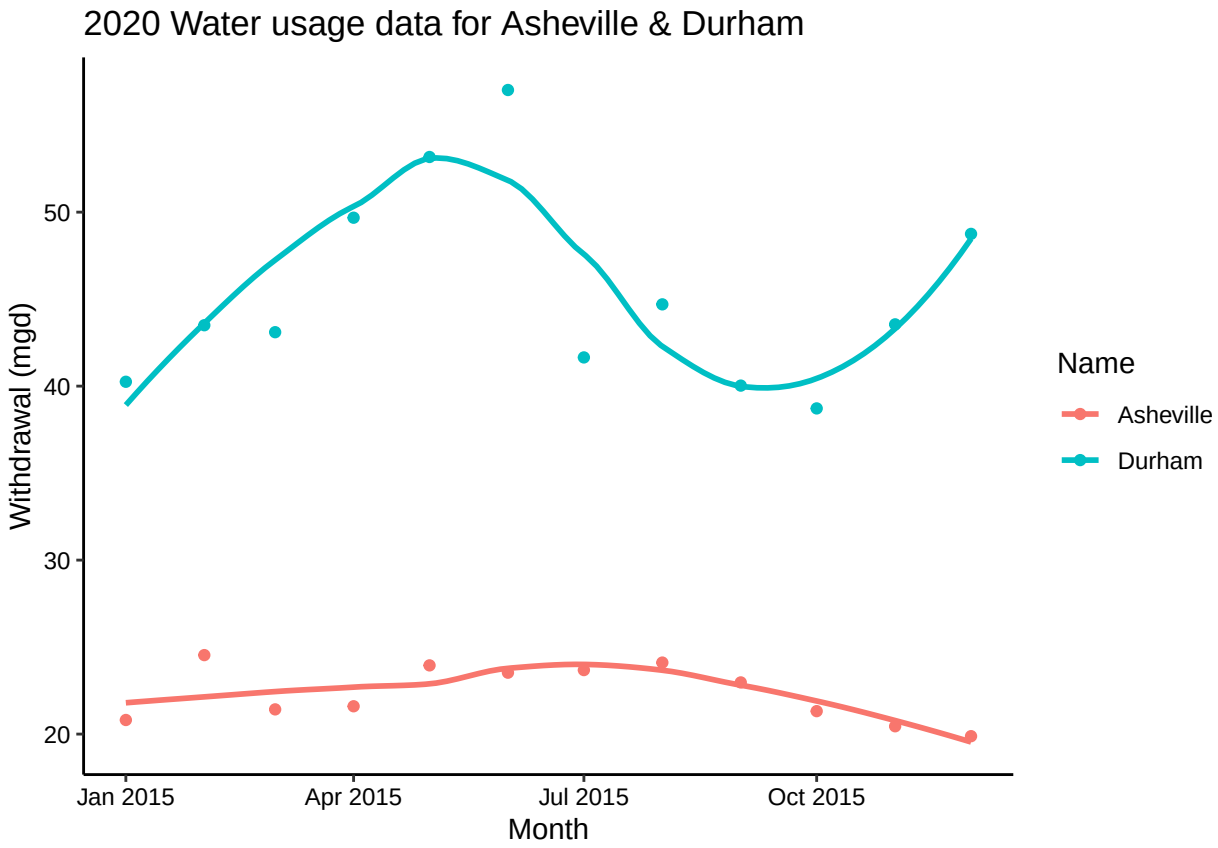
```
#7
durham.2015.df <- scrape.it(2015,"03-32-010")

ggplot(data=durham.2015.df,aes(x=Date,y=Withdrawals, group=1)) +
  geom_line(color="grey") +
  geom_point() +
  labs(title = paste("2020 Water usage data for",water.system.name),
       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 the Asheville to Durham's water withdrawals.

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

withdrawals.joined.df <- rbind(asheville.2015.df, durham.2015.df)

ggplot(data=withdrawals.joined.df,aes(x=Date,y=Withdrawals,color=Name)) +
  geom_point() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2020 Water usage data for Asheville & Durham"),
       y="Withdrawal (mgd)",
       x="Month")
```

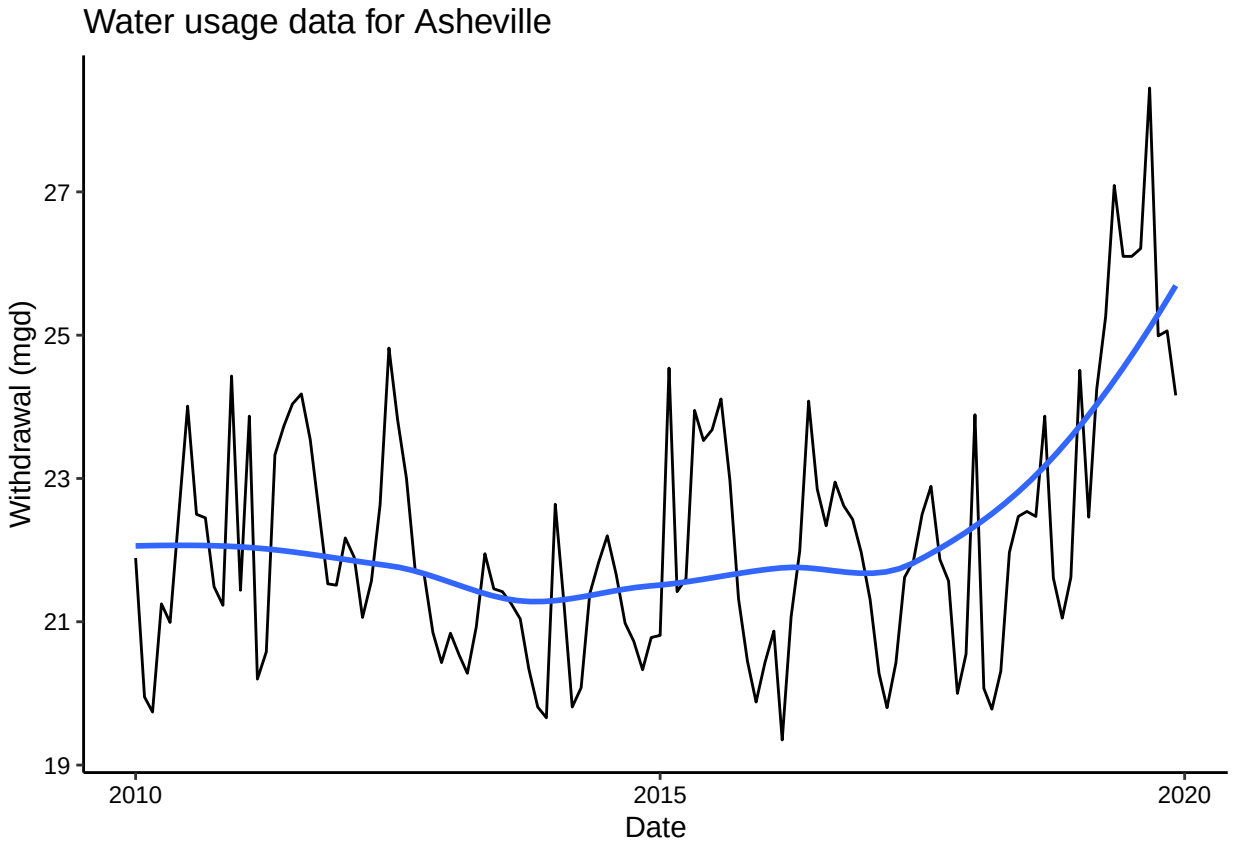
```
## `geom_smooth()` using formula 'y ~ x'
```



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 <- seq(2010, 2019)
asheville.10thr19.df <- lapply(X = the_years, "01-11-010", FUN = scrape.it) %>% bind_rows()


ggplot(data=asheville.10thr19.df,aes(x=Date,y=Withdrawals)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("Water usage data for Asheville"),
       y="Withdrawal (mgd)",
       x="Date")
```

```
## `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, it seems that after 2015 Asheville's water usage increases to the current point in 2020.