Assignment 09: Data Scraping

Blair Johnson

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:/ENV872/Environmental_Data_Analytics_2022/Assignments"
```

```
library(tidyverse)
library(rvest)
library(lubridate)
library(ggplot2)
```

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham's 2020 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
ncwater.webpage <- 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
- PWSID
- 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.

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

```
## [1] "Durham"
```

```r
pwsid <- ncwater.webpage %>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
  html_text()
pwsid
```

```
## [1] "03-32-010"
```

```r
ownership <- ncwater.webpage %>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%
html_text()
ownership
```

```
## [1] "Municipality"
```

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

```
##  [1] "36.0100" "36.9800" "41.6900" "32.0500" "40.6100" "40.5600" "37.2900"
##  [8] "43.6300" "33.3200" "32.3700" "41.9300" "28.0600"
```

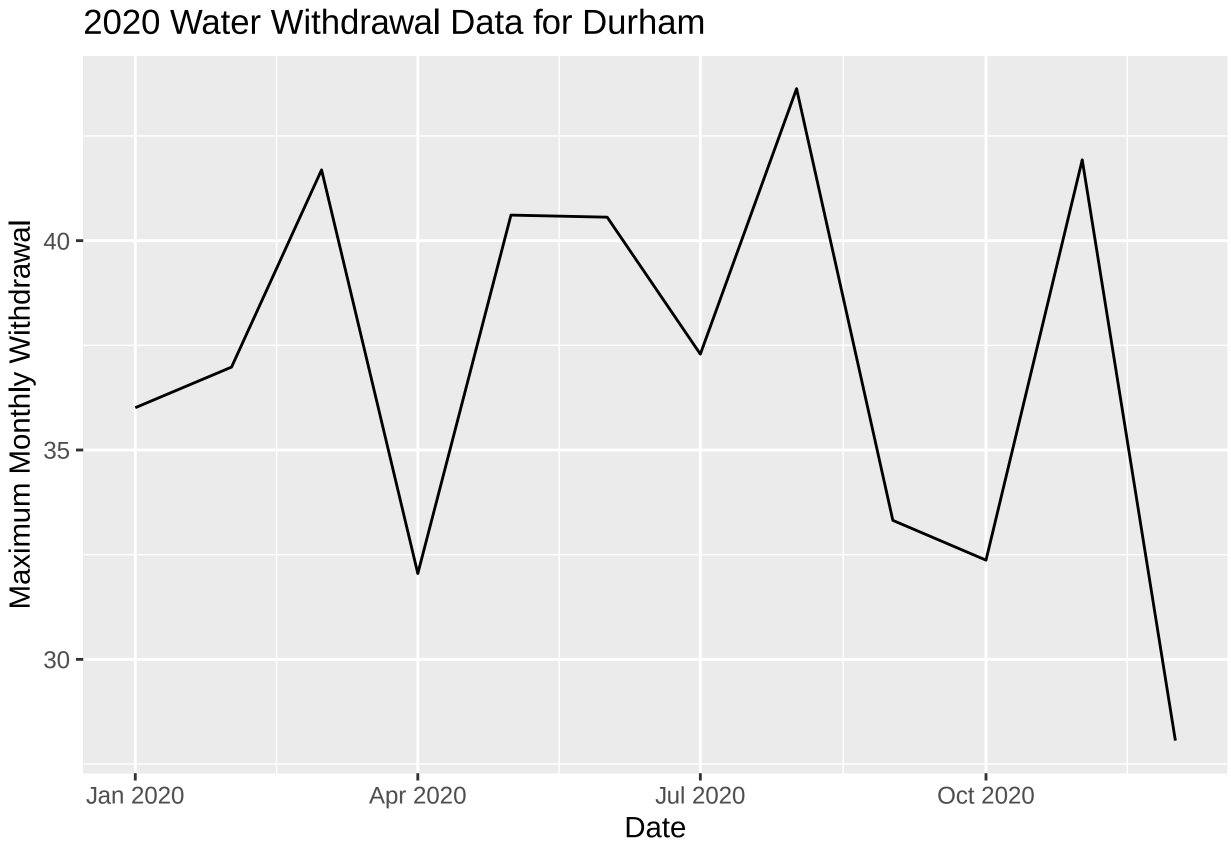
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   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
monthly.withdrawal.data <- data.frame("Month" = rep(1:12),
                                      "Year" = rep(2020, 12),
                                      "Max Monthly Withdrawals" = as.numeric(max.withdrawals.mgd))

monthly.withdrawal.data <- monthly.withdrawal.data %>%
 mutate(Ownership = !!ownership,
        PWSID = !!pwsid,
        WaterSystem = !!water.system.name,
        Date = my(paste(Month, "-", Year)))


#5
ggplot(monthly.withdrawal.data, aes(x=Date, y = Max.Monthly.Withdrawals)) +geom_line() +
  labs(y = "Maximum Monthly Withdrawal", title = paste("2020 Water Withdrawal Data for",water.system.nam
```

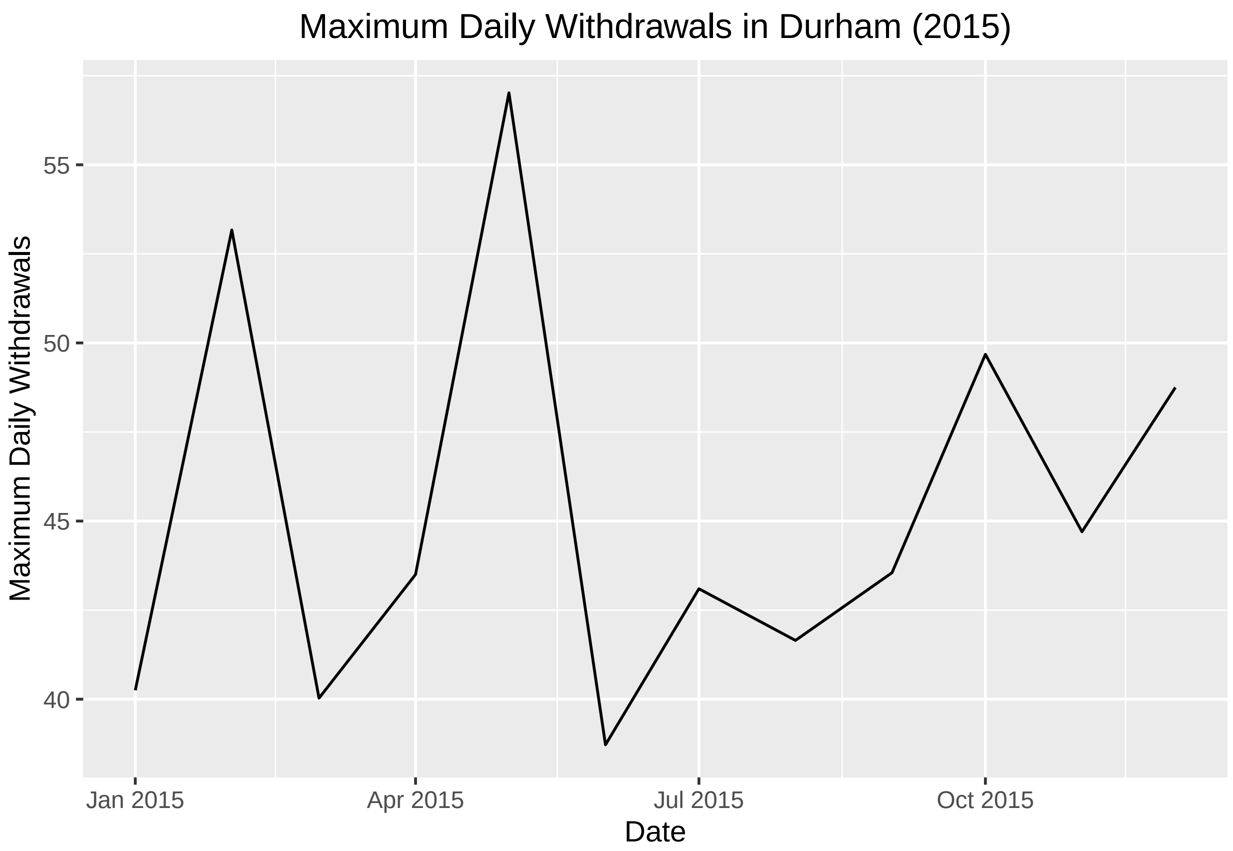


6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site scraped**.

```
#6.

scraped.data <- 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_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)'
max.withdrawals.mgd_tag <- 'th~ td+ td'

water.system.name <- the_url %>% html_nodes(water.system.name_tag) %>% html_text()
pwsid <- the_url %>% html_nodes(pwsid_tag) %>%  html_text()
ownership <- the_url %>% html_nodes(ownership_tag) %>% html_text()
max.withdrawals.mgd <- the_url %>% html_nodes(max.withdrawals.mgd_tag) %>% html_text()

scraped.df <- data.frame("Month"= rep(1:12),
                         "Year" = rep(the_year, 12),
                         "Max Monthly Withdrawals" = as.numeric(max.withdrawals.mgd)) %>%
  mutate("Water.System.Name" = as.character(water.system.name),
        "PWSID" = as.character(pwsid),
        "Ownership" = as.character(ownership),
         Date = my(paste(Month, "-", Year)))

return(scraped.df)

}
```

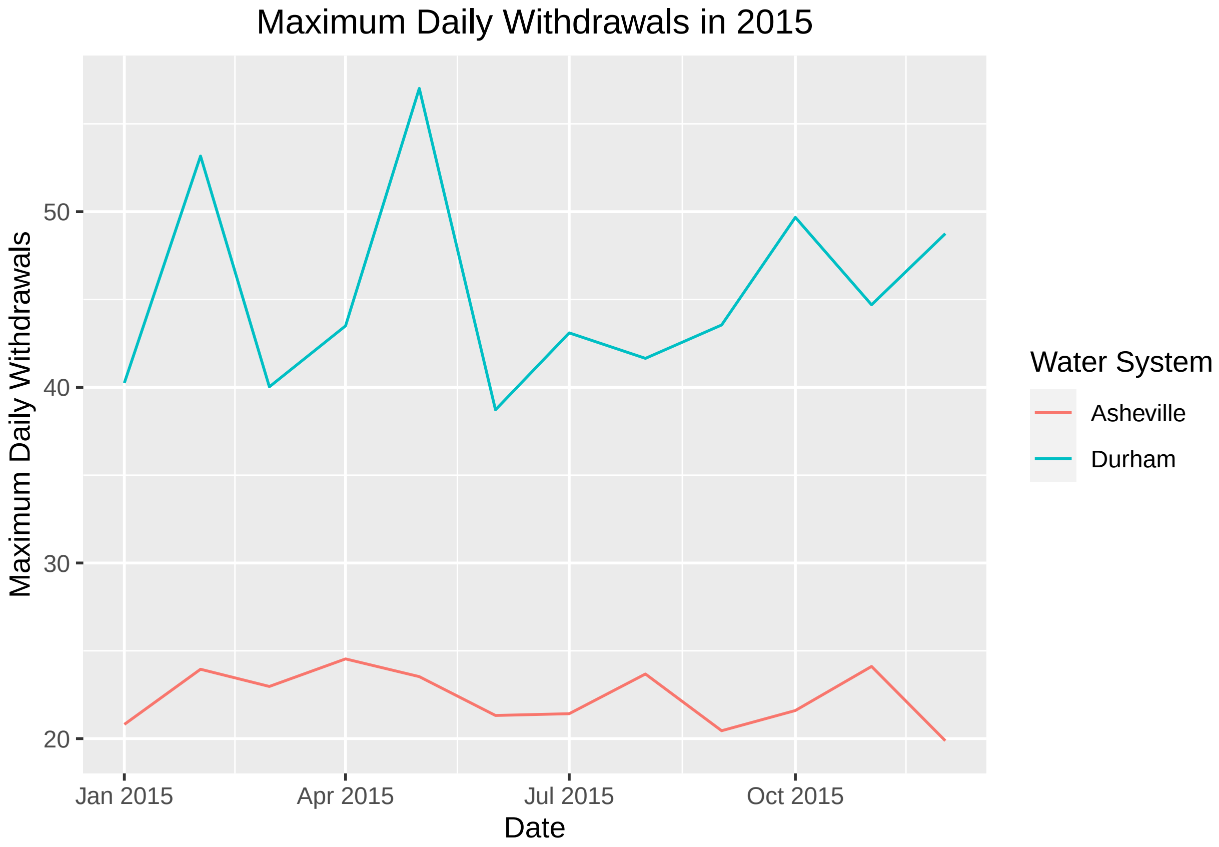
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID='03-32-010') for each month in 2015

```
#7

withdrawals.2015 <-scraped.data(2015, '03-32-010')

ggplot(withdrawals.2015, aes(x=Date, y=Max.Monthly.Withdrawals)) +geom_line() +
  labs(y= "Maximum Daily Withdrawals", title= "Maximum Daily Withdrawals in Durham (2015)") +
  theme(plot.title = element_text(hjust=0.5))
```



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.

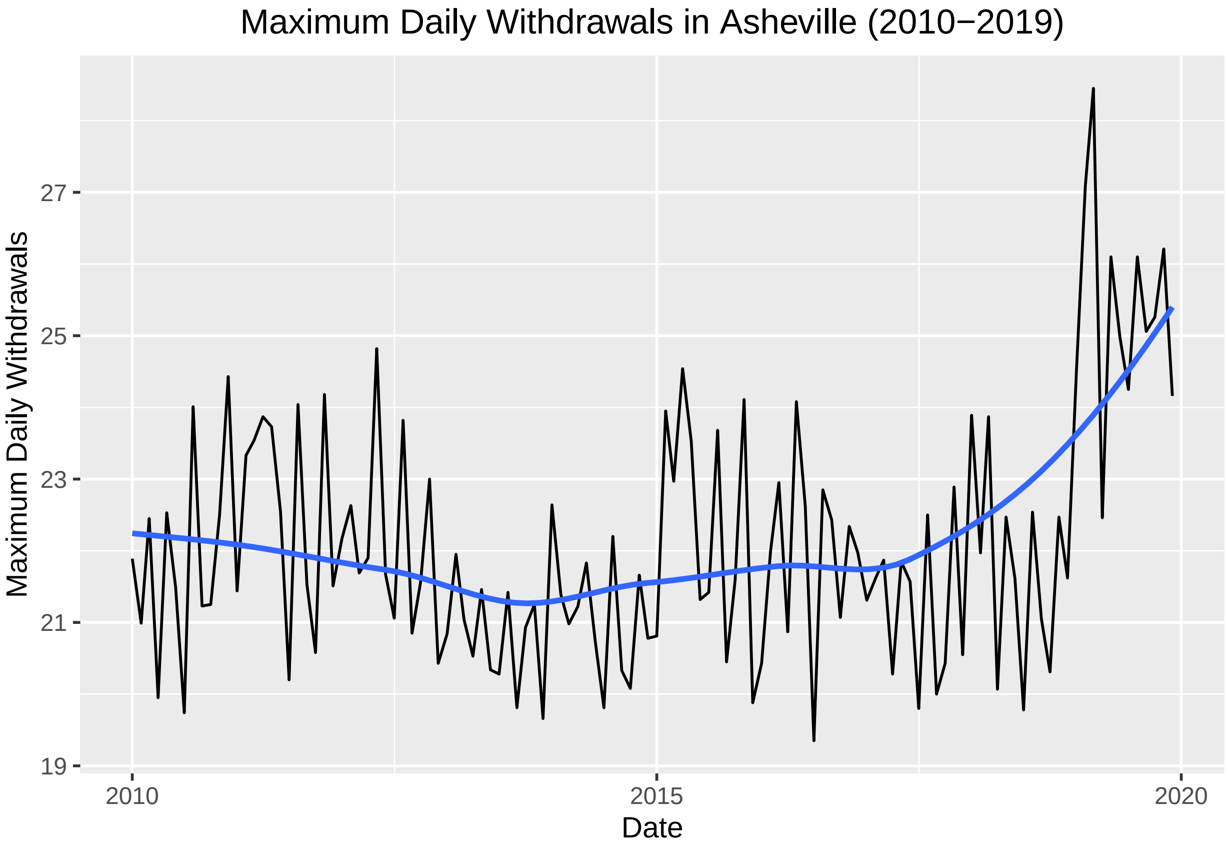
```
#8

asheville.withdrawals <-scraped.data(2015, '01-11-010')

withdrawals.combined <- rbind(asheville.withdrawals, withdrawals.2015)


ggplot(withdrawals.combined, aes(x=Date, y=Max.Monthly.Withdrawals, color=Water.System.Name))  +
geom_line() + labs(title="Maximum Daily Withdrawals in 2015") + theme(plot.title = element_text(hjust=0
```



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)
the_facility = '01-11-010'

asheville.withdrawals.2 <-map(the_years, scraped.data, the_facility)
asheville.withdrawals.2.df <-bind_rows(asheville.withdrawals.2)

ggplot(asheville.withdrawals.2.df, aes(x=Date, y=Max.Monthly.Withdrawals)) +geom_line() +
  geom_smooth(method="loess",se=FALSE) +labs(title="Maximum Daily Withdrawals in Asheville (2010-2019)"
y = "Maximum Daily Withdrawals") +theme(plot.title = element_text(hjust=0.5))
```

```
## `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 the plot, Asheville's water usage increases significantly after 2017 and peaks around 2018/2019. Therefore, the water usage increases over time.