Assignment 09: Data Scraping

Lambert Ngenzi

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/ln113/Documents/EDA_R/Environmental_Data_Analytics_2022/Assignments"
```

```
library(tidyverse)
library(lubridate)
library(viridis)

#install.packages("rvest")
library(rvest)

#install.packages("dataRetrieval")
library(dataRetrieval)

#install.packages("tidycensus")
library(tidycensus)
```

```
## Warning: package 'tidycensus' was built under R version 4.1.3
```

```
# Set theme
mytheme <- theme_classic() +
  theme(axis.text = element_text(color = "black"),
```

```
        legend.position = "top")
theme_set(mytheme)
```

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

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

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

max.withdrawals.mgd <-  Website %>%
  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.)

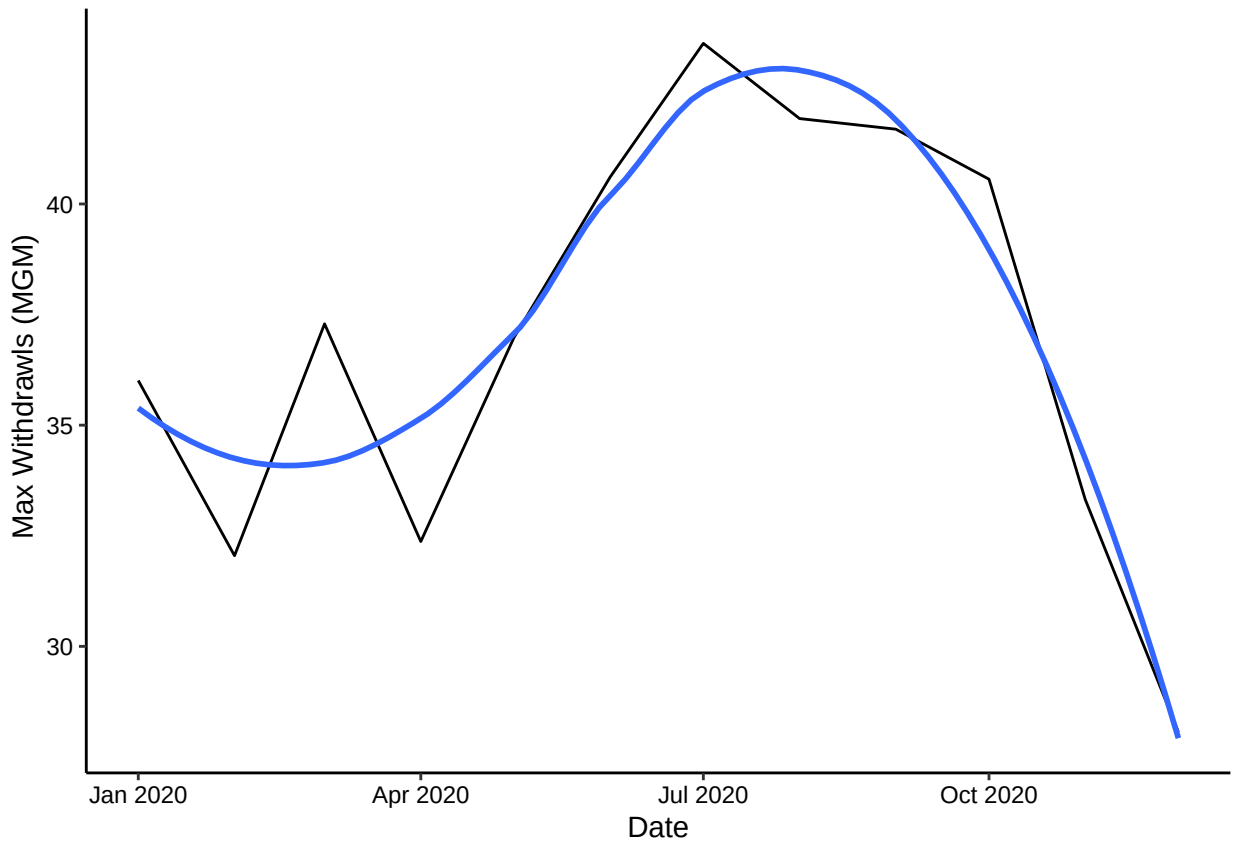
> TIP: Use `rep()` to repeat a value when creating a dataframe.

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

df_NC_Water <- data.frame("Water_System" = as.character(water.system.name),
                          "PSWID" = as.character(pswid),
                          "Ownership" = as.character(ownership),
                          "max.withdrawals" = as.numeric(max.withdrawals.mgd),
                         "Month" = c("January", "May", "September", "February", "June", "October", "Mar
                                     "July", "November", "April", "August", "December"),
                        "Year" = rep(2020, 12)) %>%
                          mutate(Date = my(paste(Month,"-",Year))) %>%
                          select("Water_System", "PSWID", "Ownership", "max.withdrawals", "Date")


#5

ggplot(df_NC_Water,aes(x=Date,y=max.withdrawals)) +
  geom_line() +
 geom_smooth(method="loess",se=FALSE) +
  labs( X = "Date", y = " Max Withdrawls (MGM)")
```

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



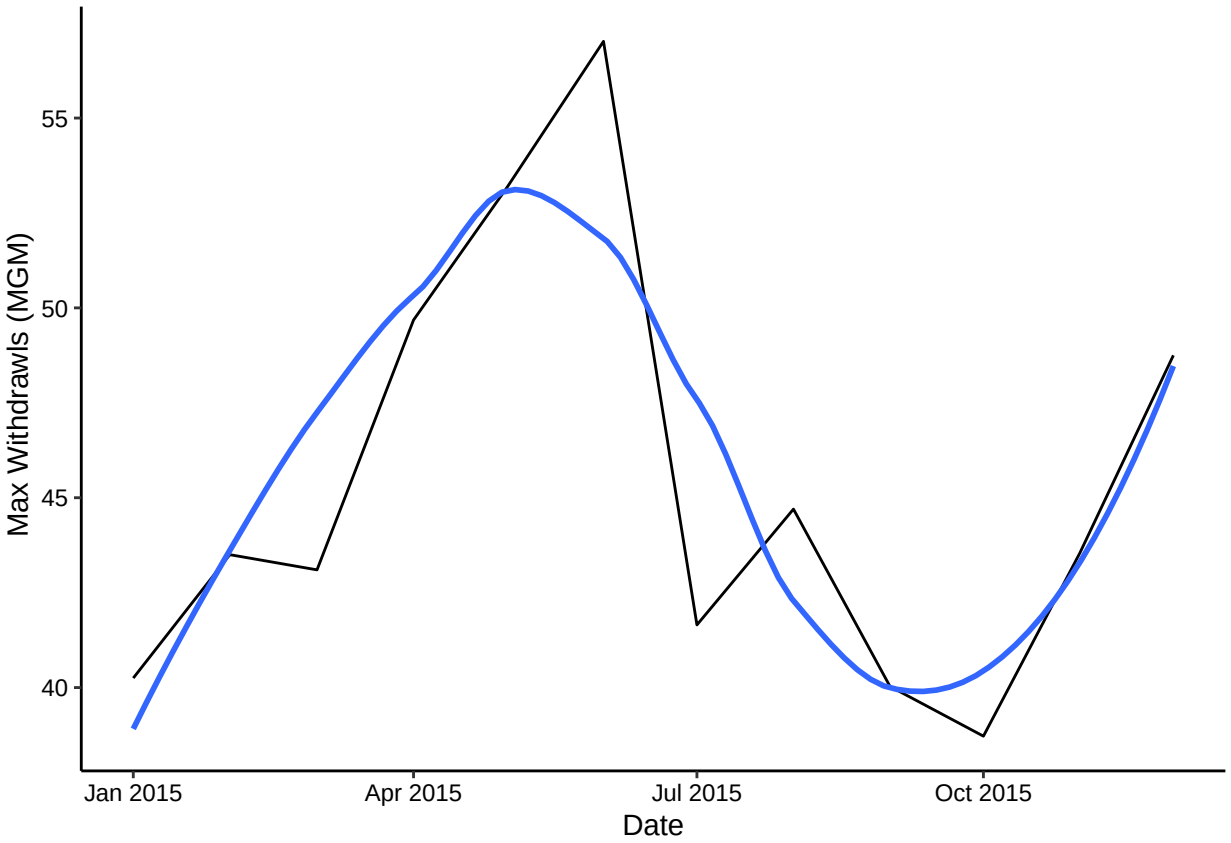
6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site scraped**.

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

URL_site <-
  read_html(paste0("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=", the_year, "&year=",the_psw

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

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

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

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

df_URL_site <- data.frame("Water_System" = water.system.name,
                          "PSWID" = pswid,
                          "Ownership" = ownership,
                          "max.withdrawals" = as.numeric(max.withdrawals.mgd),
                         "Month" = c("January", "May", "September", "February", "June", "October", "Mar
                                     "July", "November", "April", "August", "December"),
                        "Year" = rep(the_pswid, 12)) %>%
  mutate(Date = my(paste(Month,"-", Year)))

return(df_URL_site)
}
```

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

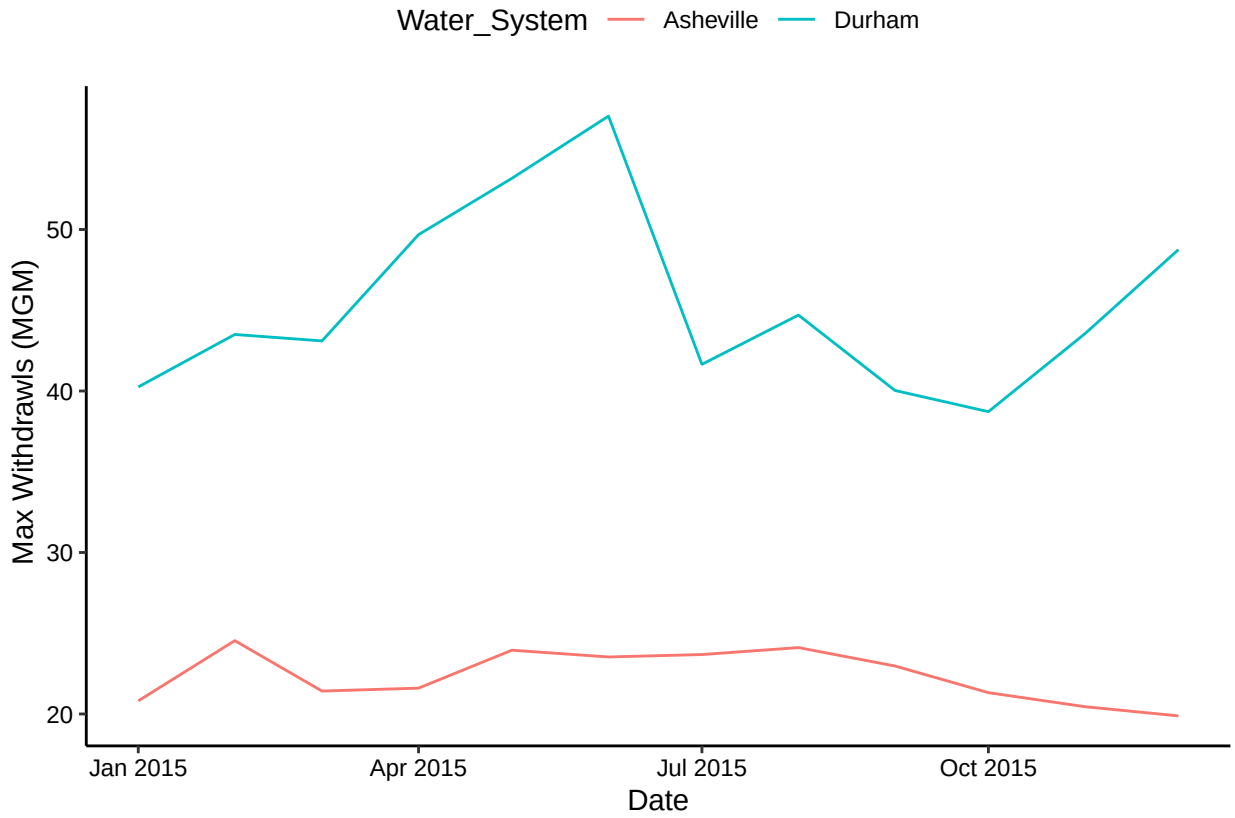
```
#7
PWSID_Durham15 <- scrape.it("03-32-010", 2015)
view(PWSID_Durham15)

ggplot(PWSID_Durham15 ,aes(x=Date,y=max.withdrawals)) +
  geom_line() +
 geom_smooth(method="loess",se=FALSE) +
  labs( X = "Date", y = " Max Withdrawls (MGM)")
```

```
## `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.

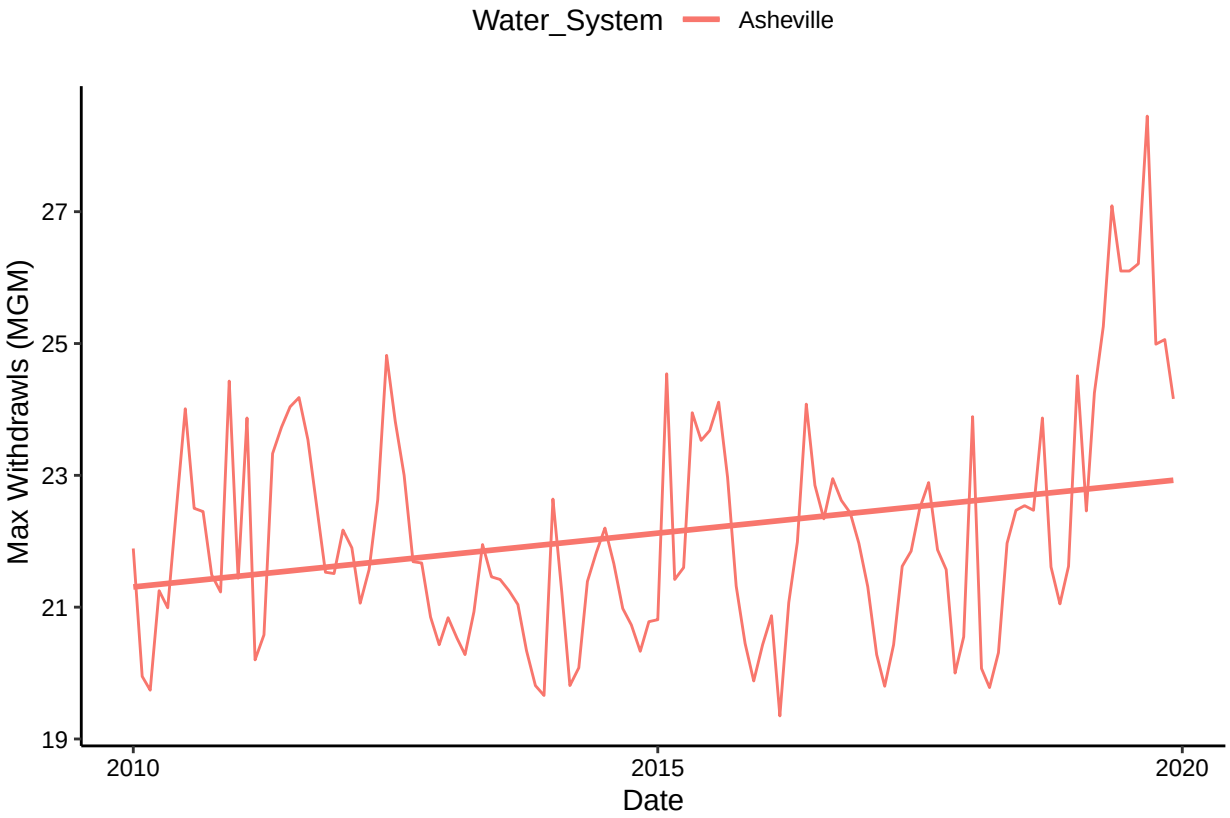
```
#8
PSWID_Asheville15 <- scrape.it("01-11-010", 2015)

Durham_Asheville15 <- rbind(PWSID_Durham15, PSWID_Asheville15)

ggplot(Durham_Asheville15,aes(x=Date,y=max.withdrawals, color = Water_System)) +
  geom_line() +
 #geom_smooth(method="loess",se=FALSE) +
  labs(X = "Date", y = " Max Withdrawls (MGM)")
```



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

Ash_PWSID <- "01-11-010"
Ash_Years <- rep(2010:2019)

Asheville_with_9y <- map2(Ash_PWSID, Ash_Years, scrape.it)

Ash_Together <- bind_rows(Asheville_with_9y)

ggplot(Ash_Together,aes(x=Date,y=max.withdrawals, color = Water_System)) +
  geom_line() +
 geom_smooth(method="lm",se=FALSE) +
  labs( X = "Date", y = " Max Withdrawls (MGM)")
```

```
## `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, Asheville water usage increases over time.