Assignment 09: Data Scraping

Isaac Benaka

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on data scraping.

Directions

1. Rename this file `<FirstLast>_A09_DataScraping.Rmd` (replacing `<FirstLast>` with your first and last name).
2. Change "Student Name" on line 3 (above) with your name.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.

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/isaacbenaka/Desktop/Fall 2022/872 - Data Analytics/EDA-Fall2022"
```

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

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham's 2021 Municipal Local Water Supply Plan (LWSP):

- Navigate to https://www.ncwater.org/WUDC/app/LWSP/search.php
- 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=2021

Indicate this website as the as the URL to be scraped. (In other words, read the contents into an `rvest` webpage object.)

```
#2
  webpage <- read_html("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2021")
```

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:
- Maximum Daily Use (MGD) - for each month

In the code chunk below scrape these values, assigning them to four 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 (represented as strings), with the first value being "27.6400".

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

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

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

max.withdrawals.mgd <- webpage %>%
  html_nodes("th~ td:nth-child(3) , th~ td+ td:nth-child(6) , th~ td:nth-child(9)") %>%
  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.)

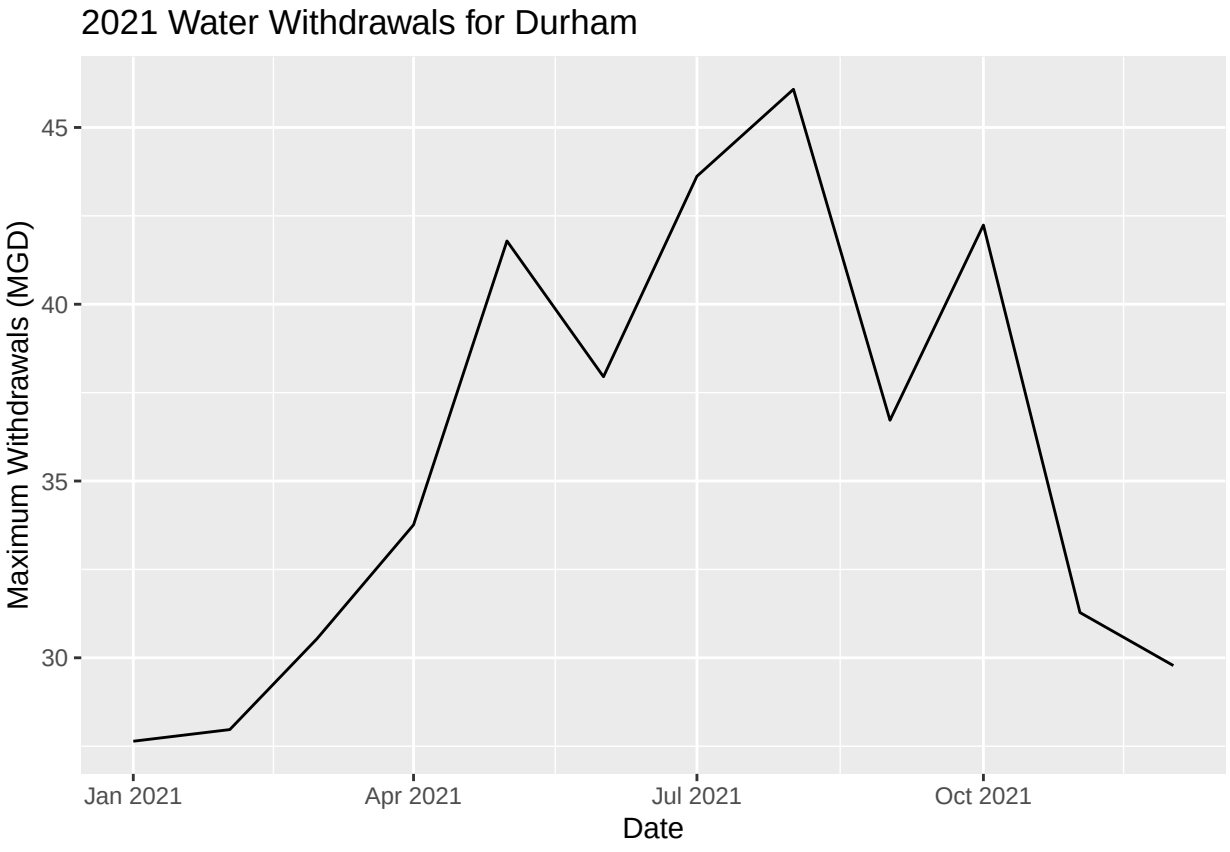
   > TIP: Use `rep()` to repeat a value when creating a dataframe.
   >
   > NOTE: It's likely you won't be able to scrape the monthly widthrawal data in chronological order. You can overcome this by creating a month column manually assigning values in the order the data are scraped: "Jan", "May", "Sept",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

5. Create a line plot of the maximum daily withdrawals across the months for 2021

```r
#4
df_withdrawals <- data.frame(
                              "year" = rep(2021),
                              "Water_System_Name" = rep(water.system.name),
                              "PSWID" = rep(pswid),
                              "Ownership" = rep(ownership),
                              "Max_Withdrawals_mgd" = as.numeric(max.withdrawals.mgd))
df_withdrawals$Month <- c("1-1-2021","5-1-2021","9-1-2021","2-1-2021","6-1-2021","10-1-2021","3-1-2021"
df_withdrawals$Month <- as.Date(df_withdrawals$Month, format="%m-%d-%Y")
df_withdrawals  <- df_withdrawals %>% arrange(mdy(df_withdrawals$Month))
```

```
## Warning: All formats failed to parse. No formats found.
```

```r
df_withdrawals <- df_withdrawals %>% rename(Date=Month)


#5
ggplot(df_withdrawals,aes(x=Date,y=Max_Withdrawals_mgd))+
  geom_line()+
  ylab("Maximum Withdrawals (MGD)")+
  labs(title = paste("2021 Water Withdrawals for",water.system.name))
```

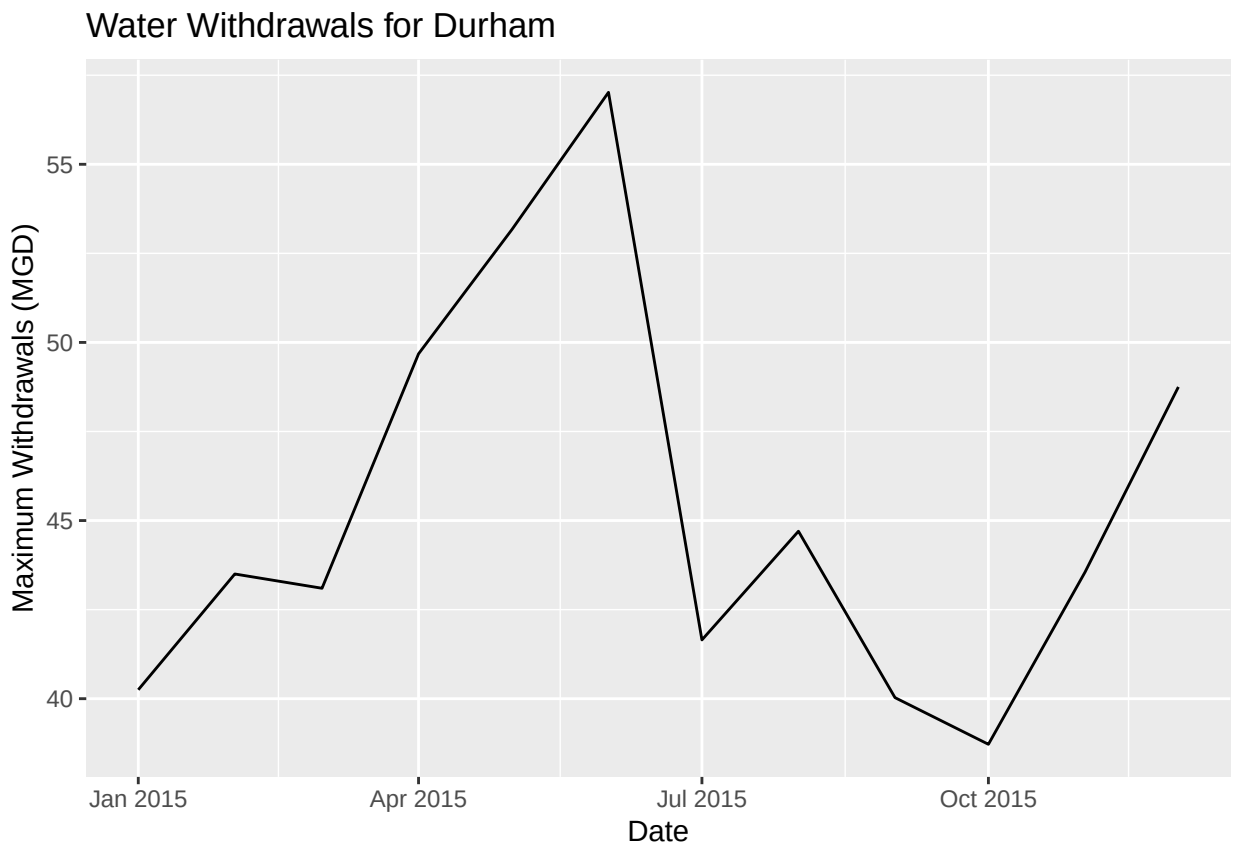


6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

```
#6.
scrape.it <- function(year, PSW.ID){

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

  water.system.tag <- "div+ table tr:nth-child(1) td:nth-child(2)"
  PSW.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:nth-child(3) , th~ td+ td:nth-child(6) , th~ td:nth-child(9)"

  Water.sys.name <- theURL %>% html_nodes(water.system.tag) %>% html_text()
  PSW <- theURL %>% html_nodes(PSW.tag) %>% html_text()
  Owner <- theURL %>% html_nodes(ownership.tag) %>% html_text()
  MGD <- theURL %>% html_nodes(max.withdrawals.mgd.tag) %>% html_text()

   Monthnum=c(1,5,9,2,6,10,3,7,11,4,8,12)

  df_withdrawals_scrape <- data.frame(
    Year=rep(year),
    Date = make_date(year=year, month=Monthnum, day=1),
    Water.Sys = rep(Water.sys.name),
    PSWID = rep(PSW),
    Owner = rep(Owner),
    withdrawal.mgd = as.numeric(MGD))

}
```

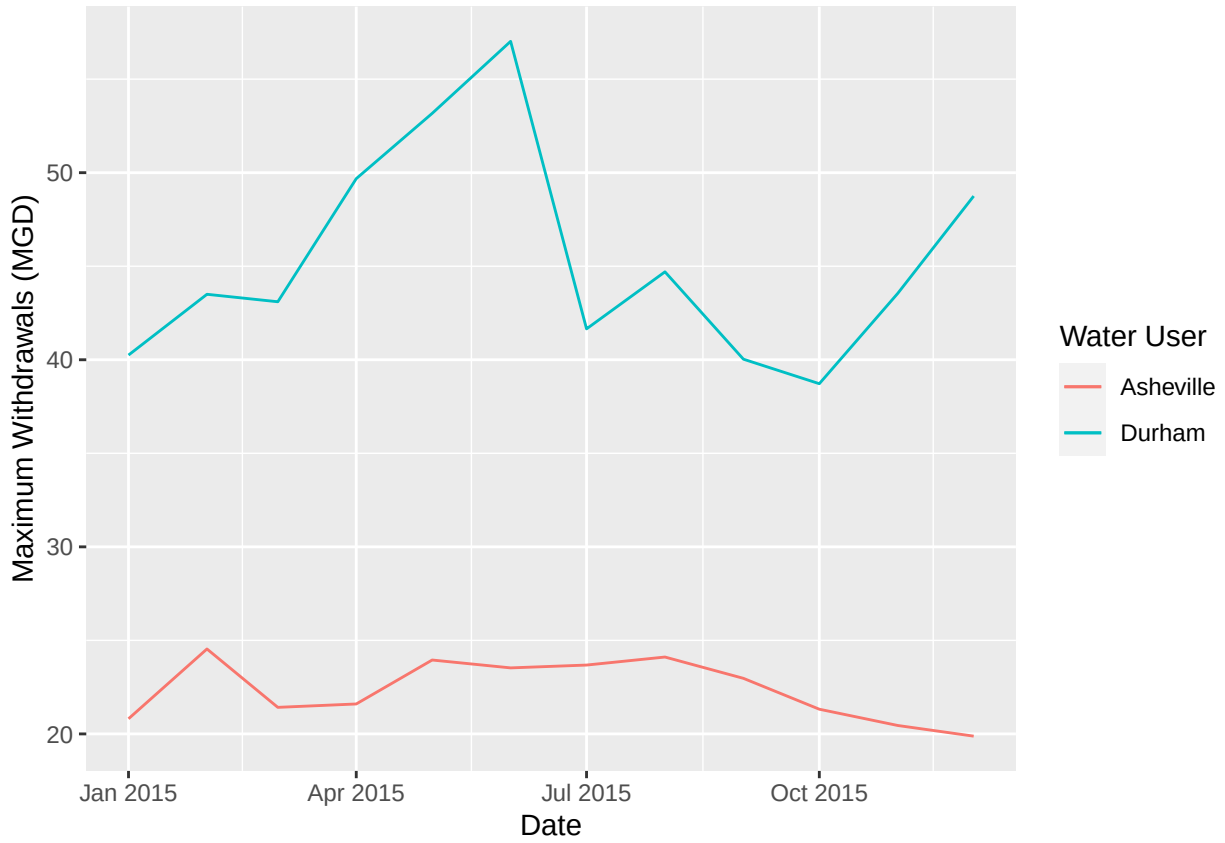
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID='03-32-010') for each month in 2015

```
#7
Durham2015 <- scrape.it(2015,"03-32-010")
view(Durham2015)

ggplot(Durham2015,aes(x=Date,y=withdrawal.mgd))+
  geom_line()+
  ylab("Maximum Withdrawals (MGD)")+
  labs(title = paste("Water Withdrawals for Durham"))
```



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 Asheville's to Durham's water withdrawals.

```
#8
Asheville2015 <- scrape.it(2015,"01-11-010")
DurhAshe2015 <- inner_join(Durham2015,Asheville2015,by="Date")


ggplot(DurhAshe2015)+
  geom_line(aes(x=Date,y=withdrawal.mgd.x, color=Water.Sys.x))+
  geom_line(aes(x=Date,y=withdrawal.mgd.y, color=Water.Sys.y))+
  ylab("Maximum Withdrawals (MGD)")+
  labs(color = "Water User")
```

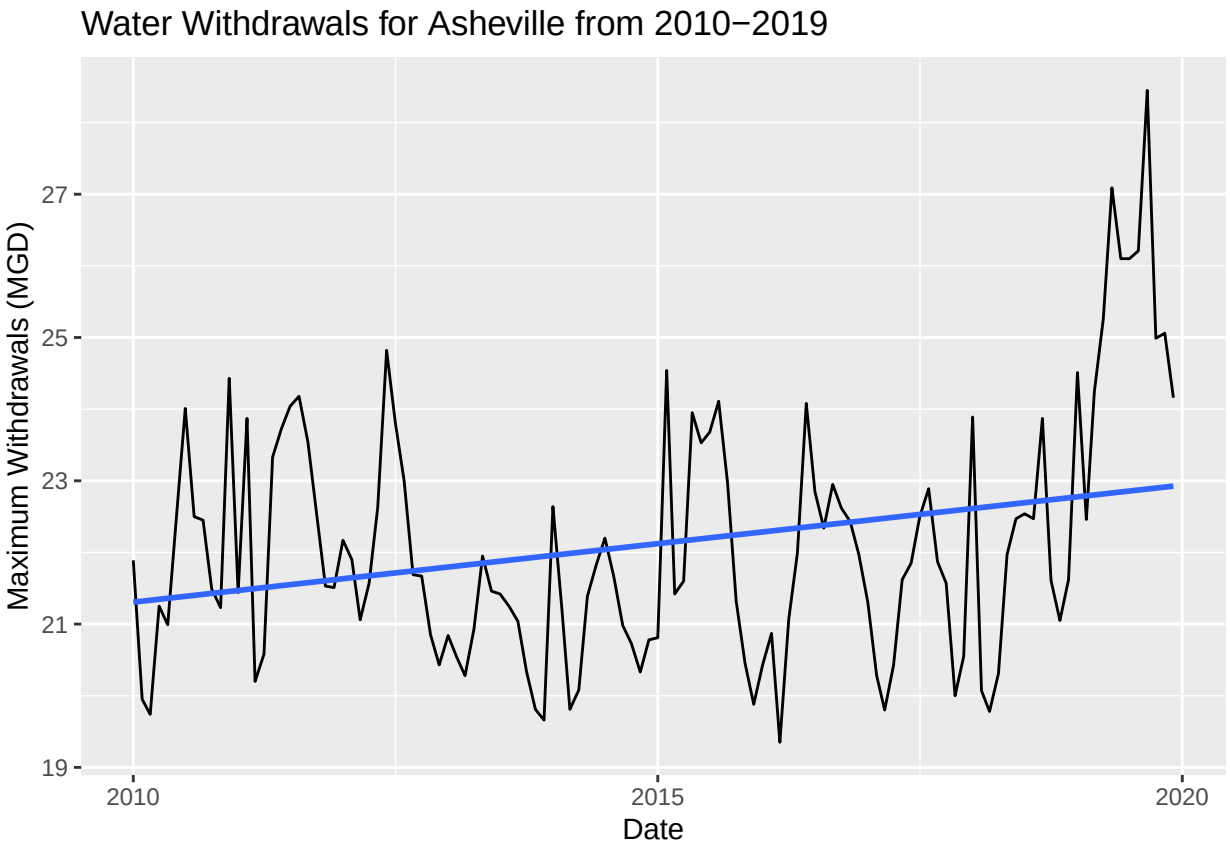


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

   TIP: See Section 3.2 in the "09_Data_Scraping.Rmd" where we apply "map2()" to iteratively run a function over two inputs. Pipe the output of the map2() function to `bindrows()` to combine the dataframes into a single one.

```
#9
the_years = rep(2010:2019)
location = "01-11-010"

Asheville20102019 <- map(the_years,scrape.it,PSW.ID=location)

Asheville2010s <- bind_rows(Asheville20102019)

ggplot(Asheville2010s,aes(x=Date,y=withdrawal.mgd))+
  geom_line()+
  ylab("Maximum Withdrawals (MGD)")+
  labs(title = paste("Water Withdrawals for Asheville from 2010-2019"))+
  geom_smooth(method=lm, se=FALSE)
```

```
## `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?: **Asheville's maximum water usage has increased over time, with a steep increase since around 2018-2019.**