Assignment 10: Data Scraping

Tasneem Ahsanullah

OVERVIEW

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

Directions

1. Rename this file `<FirstLast>_A10_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 your code is tidy; use line breaks to ensure your code fits in the knitted output.
5. Be sure to **answer the questions** in this assignment document.
6. When you have completed the assignment, **Knit** the text and code into a single PDF file.

Set up

1. Set up your session:

- Load the packages `tidyverse`, `rvest`, and any others you end up using.
- Check your working directory

```
#1
#install packages
library(tidyverse)
library(lubridate)
library(here); here()
```

```
## [1] "/Users/tasneemahsanullah/Desktop/Classes/EDA/DataAnalytics"
```

```
library(rvest)
library(dplyr)

getwd()
```

```
## [1] "/Users/tasneemahsanullah/Desktop/Classes/EDA/DataAnalytics"
```

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham's 2022 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=2022

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

```r
#2
#setting the Durham LWSP URL as object
the_website <- read_html('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2022')
```

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:
- Maximum Day 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".

```r
#3
#scraping the water system name, PWSID, ownership and max withdrawls
#for each month and setting them to variables
water.system.name <- the_website %>%
  html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>%
  html_text()

PWSID <- the_website %>%
  html_nodes('td tr:nth-child(1) td:nth-child(5)') %>%
  html_text()

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

max.withdrawals.mgd <- the_website %>%
  html_nodes('th~ td+ td') %>%
  html_text()
```

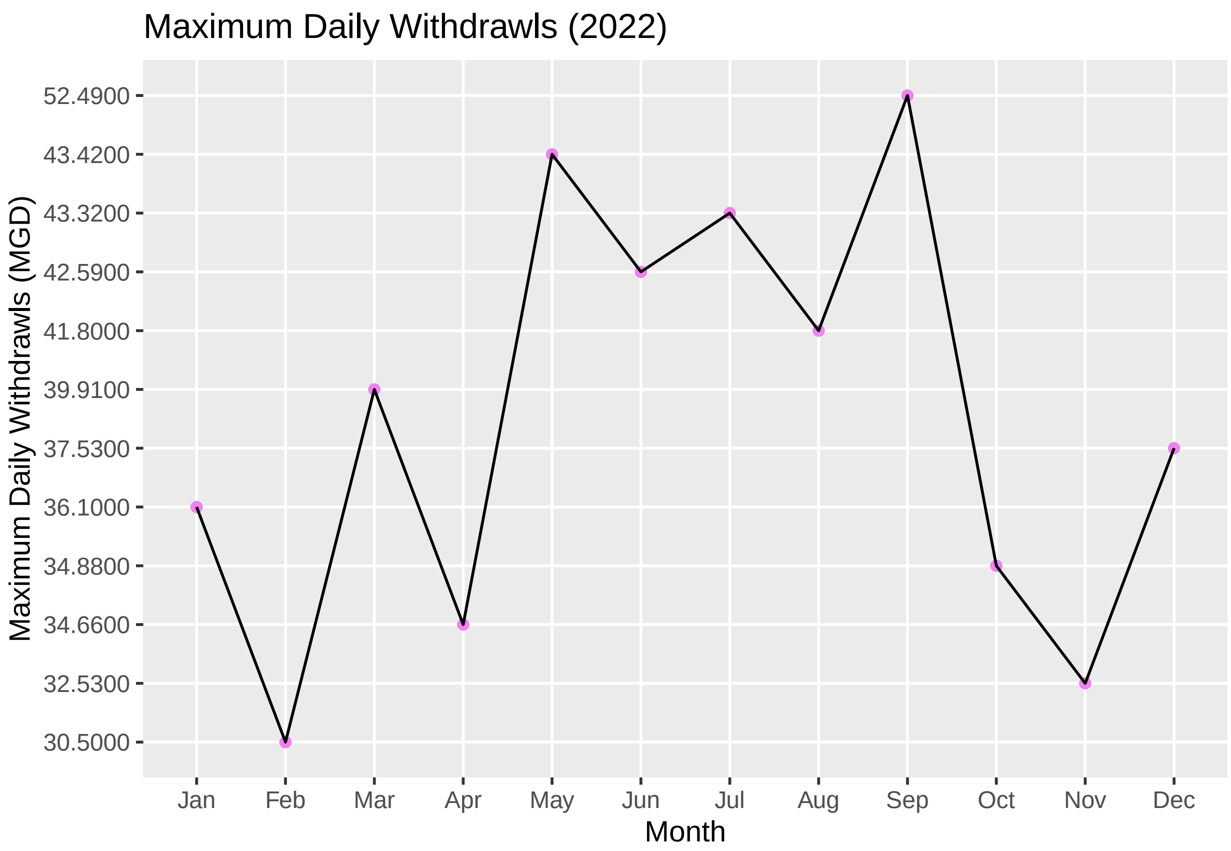
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

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 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… Or, you could scrape month values from the web page…

5. Create a line plot of the average daily withdrawals across the months for 2022

```
#4
#converting the scraped data into a dataframe, added month and year column
the_df <- data.frame(
  "Water System Name" = as.factor(water.system.name),
  "PWSID" = as.factor(PWSID),
  "Ownership" = as.factor(ownership),
  "Maximum Day Use(MGD)" = as.numeric(max.withdrawals.mgd),
  "Month" = c("Jan","May","Sep","Feb","Jun","Oct","Mar","Jul","Nov","Apr","Aug","Dec"),
  "Year" =rep(2022,12)
)

#add date column
the_df <- the_df %>%
  mutate(Date = my(paste(Month,"-",Year)))

#5
#line plot of max daily withdrawals across months for 2022
ggplot(the_df,aes(x=factor(Month,
      level=c("Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct","Nov","Dec")), y=max.withdrawa
  geom_point(colour="violet") +
  geom_line(colour="black",group=1) +
  labs(x = "Month",y = "Maximum Daily Withdrawls (MGD)",title="Maximum Daily Withdrawls (2022)")
```

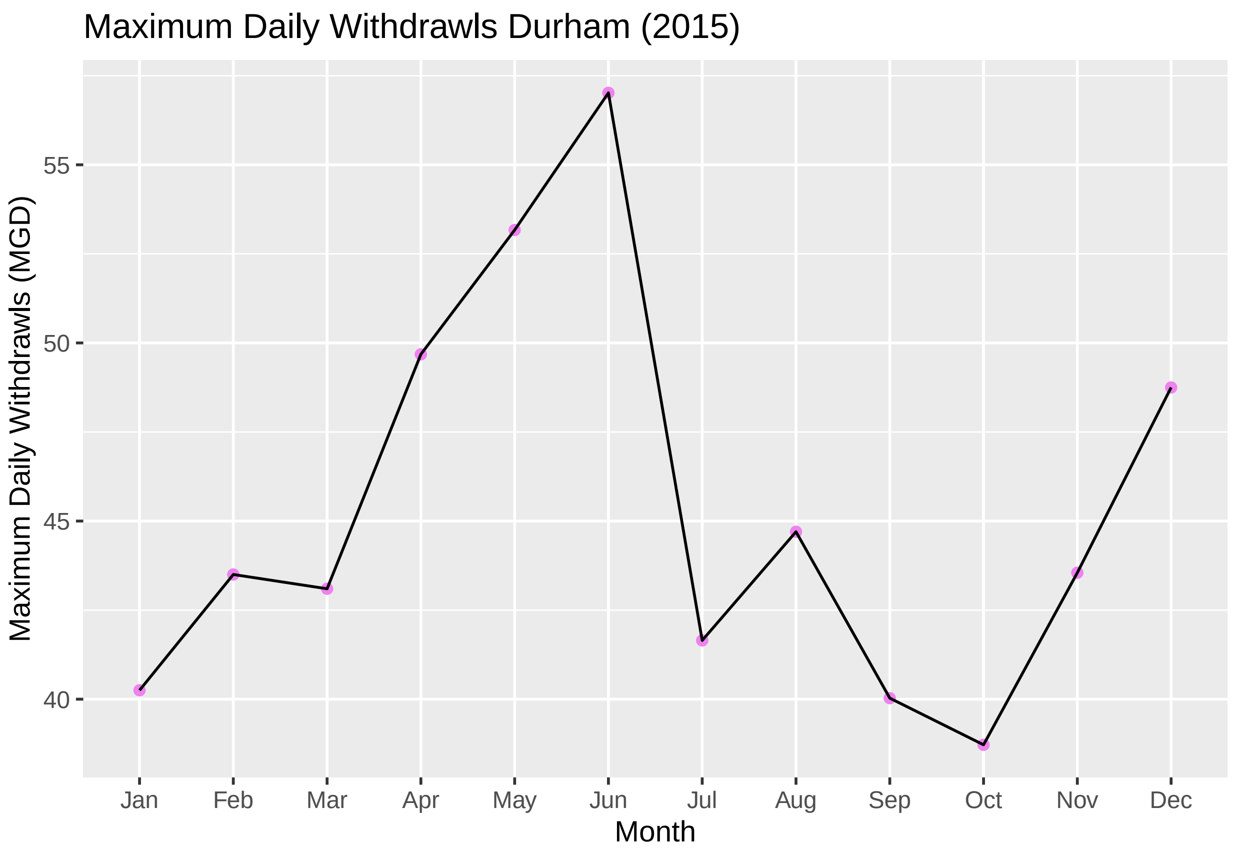


6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

```
#6.
#creating scrape it function for any year or PWSID
scrape.it <- function(the_year, PWSID){

#Retrieving the website contents so year and PWISD can be changeable
  the_website <- read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',
    PWSID,'&year=',the_year))


#scraping the water system name, PWSID, ownership and max withdrawls
#for each month and setting them to variables
water.system.name <- the_website %>%
  html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>%
  html_text()

PWSID <- the_website %>%
  html_nodes('td tr:nth-child(1) td:nth-child(5)') %>%
  html_text()

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

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

#converting the scraped data into a dataframe, added month and year column
  the_df <- data.frame(
  "Water System Name" = as.factor(water.system.name),
  "PWSID" = as.factor(PWSID),
  "Ownership" = as.factor(ownership),
  "Maximum Day Use(MGD)" = as.numeric(max.withdrawals.mgd),
  "Month" = c("Jan","May","Sep","Feb","Jun","Oct","Mar","Jul","Nov","Apr","Aug","Dec"),
  "Year" = rep(the_year,12)
)

#add date column
the_df <- the_df %>%
  mutate(Date = my(paste(Month,"-",Year)))

#Returning the dataframe
  return(the_df)
}
```

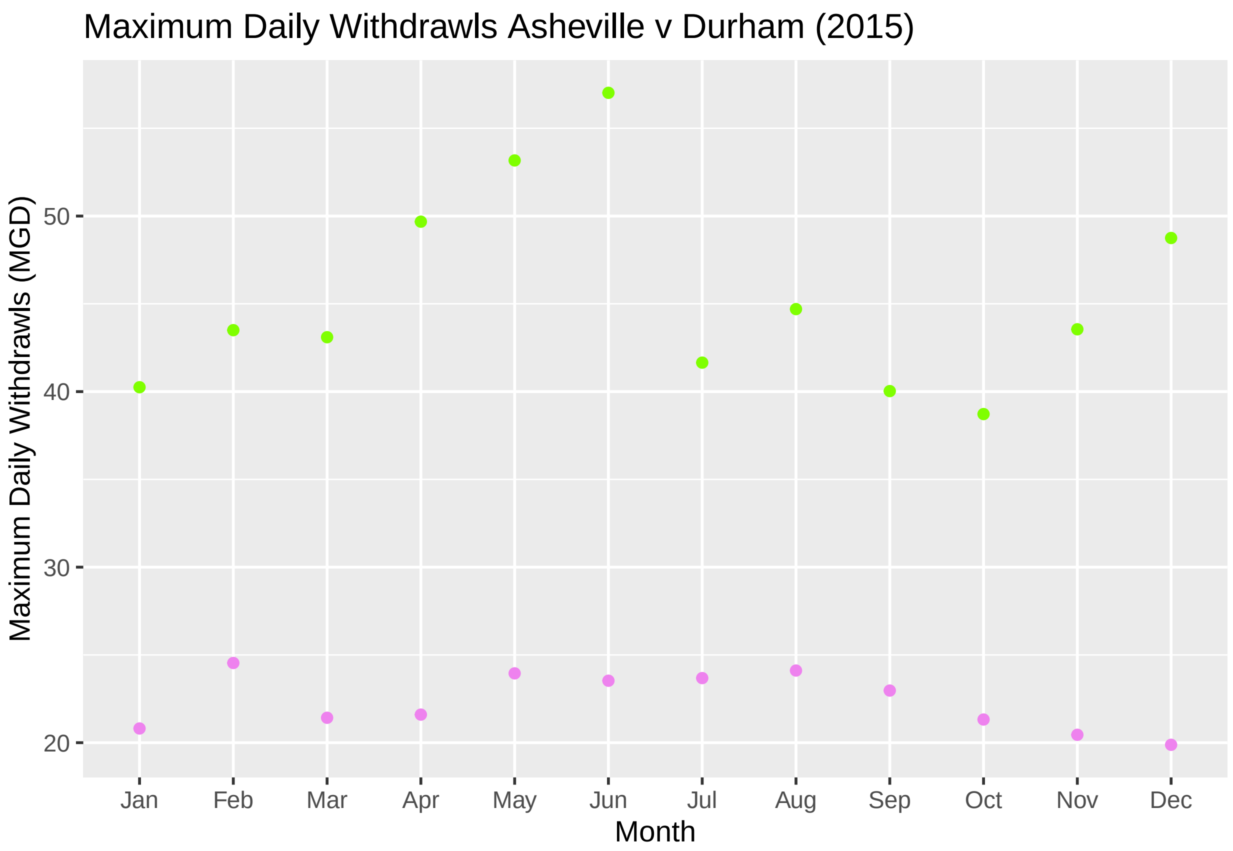
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID='03-32-010') for each month in 2015

```
#7
#calling scrape.it to extract and plot max daily withdrawls for Durham for each month in 2015
df2015 <- scrape.it(2015,"03-32-010")

ggplot(df2015,aes(x=factor(Month,
      level=c("Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct","Nov","Dec")),
      y=Maximum.Day.Use.MGD.)) +
  geom_point(colour="violet") +
  geom_line(colour="black", group=1) +
  labs(x = "Month",y = "Maximum Daily Withdrawls (MGD)",title="Maximum Daily Withdrawls Durham (2015)")
```



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
#calling scrape.it to extract and plot max daily withdrawls for Asheville for each month in 2015
df2015ash <- scrape.it(2015,"01-11-010")

ggplot()+
  geom_point(data=df2015ash,mapping=aes(x=factor(Month,
      y=Maximum.Day.Use.MGD.), color="violet") +
    geom_point(data=df2015,mapping=aes(x=factor(Month,
   level=c("Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct","Nov","Dec")),
      y=Maximum.Day.Use.MGD.), color="chartreuse") +
  labs(x = "Month",y = "Maximum Daily Withdrawls (MGD)",
       title="Maximum Daily Withdrawls Asheville v Durham (2015)")
```

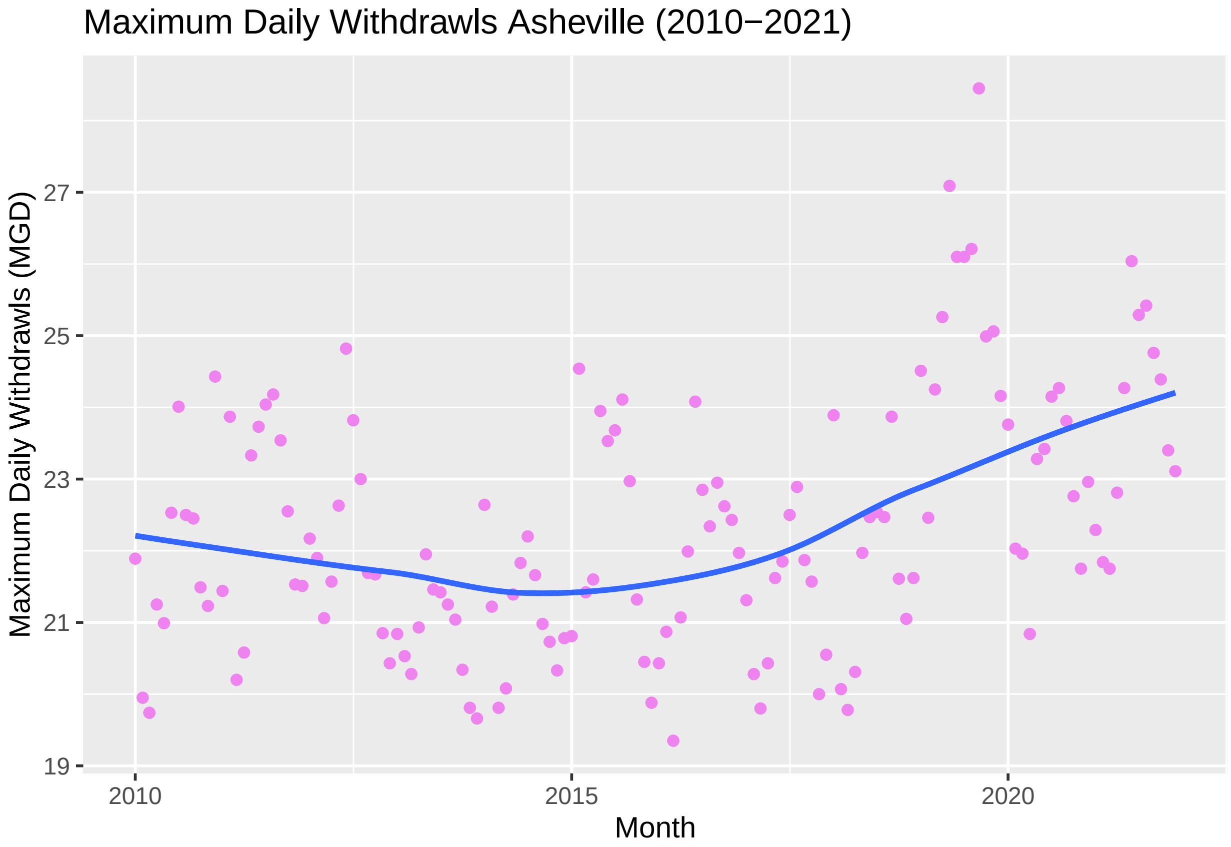


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2021.Add a smoothed line to the plot (method = 'loess').

   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
#mapping the scrape.it function to retrieve data from 2010-2021
the_years <- seq(2010,2021)
PWSID <- '01-11-010'
dfash10_21<- map2(the_years, PWSID, scrape.it)
dfash_final <-  bind_rows(dfash10_21)



#plotting Asheville's max daily withdrawal by months for the years 2010 through 2021
ggplot(dfash_final,aes(x=Date,
        y=Maximum.Day.Use.MGD.)) +
  geom_point(colour="violet") +
  geom_smooth(method = 'loess',se=F) +
  labs(x = "Month",y = "Maximum Daily Withdrawls (MGD)",
       title="Maximum Daily Withdrawls Asheville (2010-2021)")
```

```
## `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 has an increasing trend in water usage over time.