Assignment 09: Data Scraping

Azura Liu

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/yliua/Desktop/Environmental_Data_Analytics_2022/Assignments"
```

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

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 2020 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
web<-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:
- max 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 <- web %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()

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

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

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

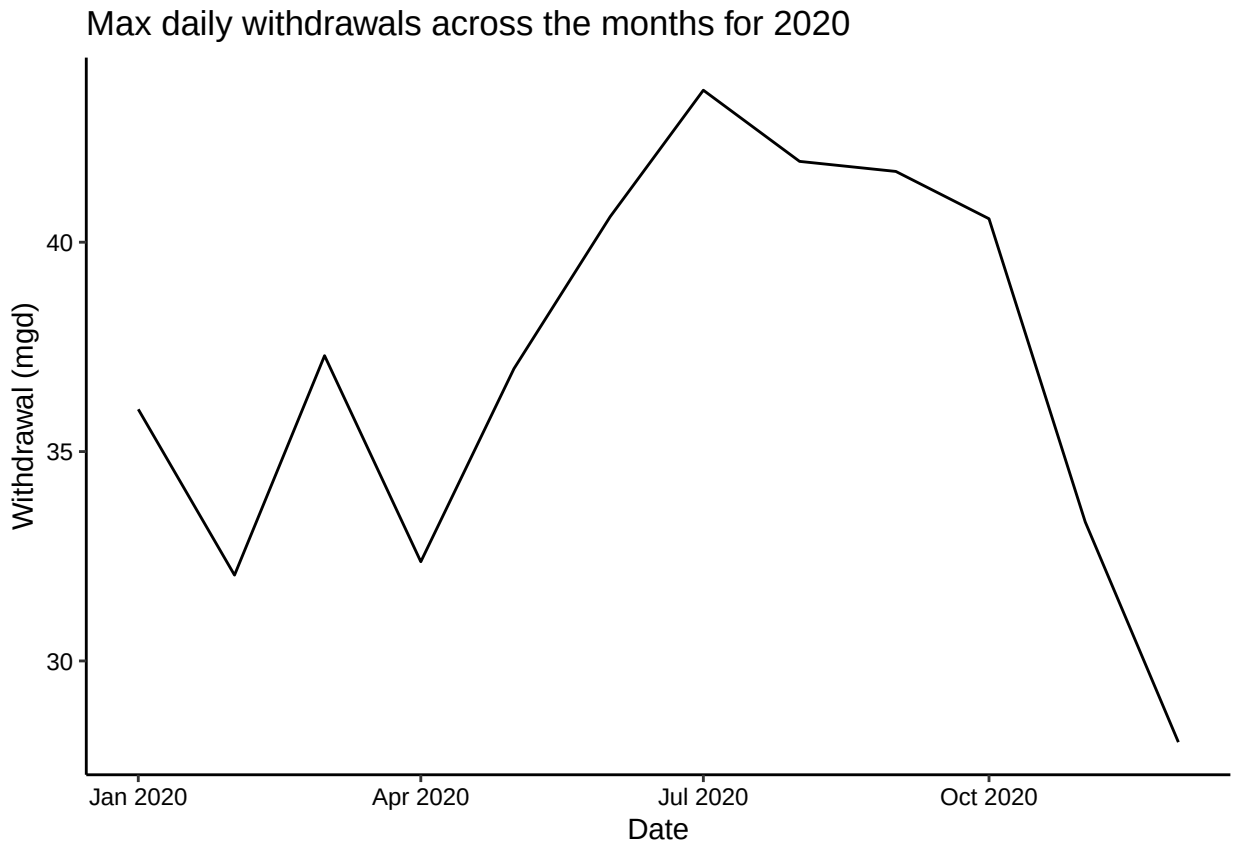
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   > 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_withdrawals <- data.frame("Month" = c("Jan","May","Sep","Feb","Jun", "Oct", "Mar","Jul","Nov","Apr",
                             "Year" = rep(2020,12),
                             "water.system.name" = water.system.name,
                             "pswid"= pswid,
                             "ownership"=ownership,
                             "max.withdrawals.mgd" = as.numeric(max.withdrawals.mgd))%>%
```

```
  mutate(Date =my(paste(Month,Year)))


#5
ggplot(df_withdrawals,aes(x= Date,y=max.withdrawals.mgd)) +
  geom_line() +
  labs(title = "Max daily withdrawals across the months for 2020",
       y="Withdrawal (mgd)",
       x="Date")
```



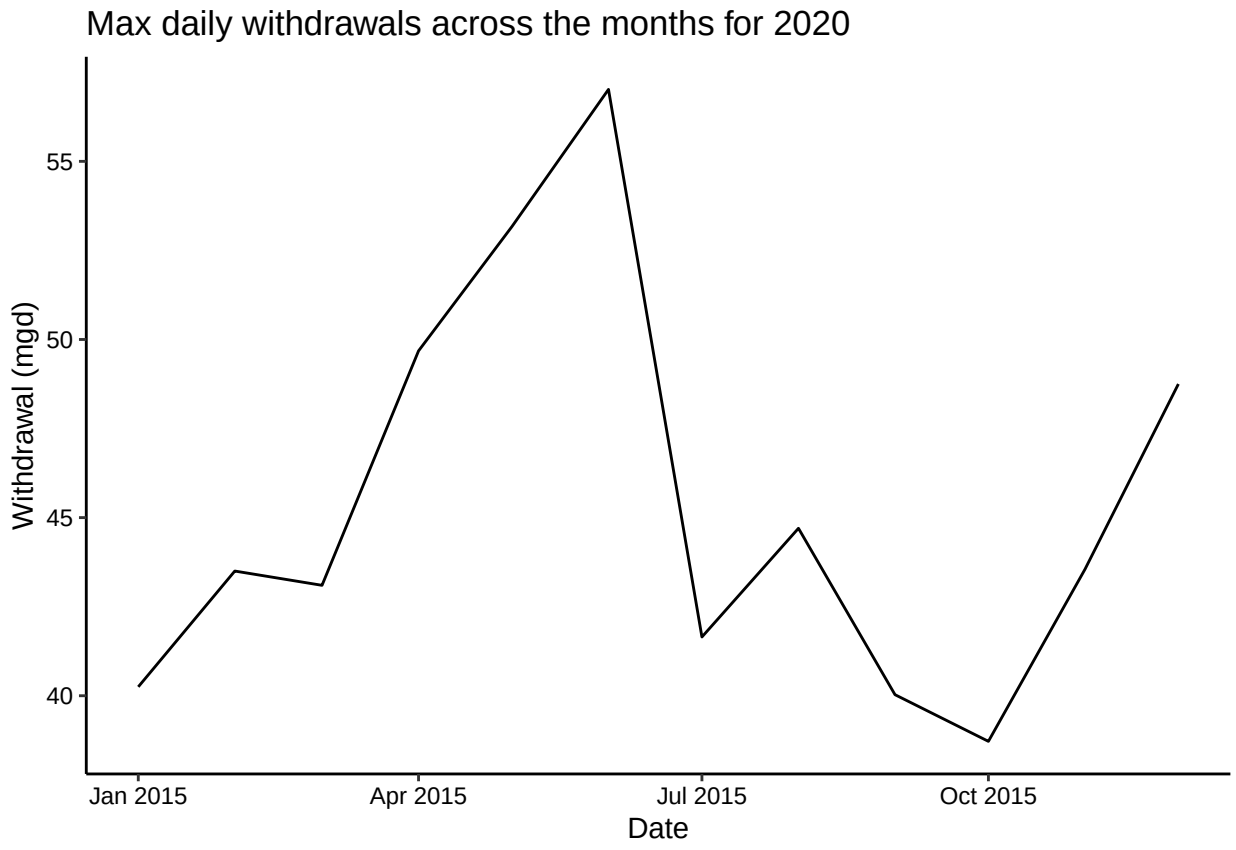
6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site scraped**.

```
#6.
scrape.it <- function(the_year, the_pwsid){
  the_scrape_web<-read_html(paste0("https://www.ncwater.org/WUDC/app/LWSP/report.php?", "pwsid=", the_pw

  water.system.name.tag <-"div+ table tr:nth-child(1) td:nth-child(2)"
  pswid.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"

  the.water.system.name <- the_scrape_web %>% html_nodes(water.system.name.tag) %>% html_text()
  the.pswid <- the_scrape_web %>%   html_nodes(pswid.tag) %>%  html_text()
  the.ownership <- the_scrape_web %>% html_nodes(ownership.tag) %>% html_text()
  the.max.withdrawals.mgd <- the_scrape_web %>% html_nodes(max.withdrawals.mgd.tag) %>% html_text()
```

```
  df_custom <- data.frame("Month" = c("Jan","May","Sep","Feb","Jun", "Oct", "Mar","Jul","Nov","Apr","Aug
                          "Year" = rep(the_year,12),
                          "max.withdrawals.mgd" = as.numeric(the.max.withdrawals.mgd))%>%
    mutate(the.water.system.name = !!the.water.system.name,
           the.pswid = !!the.pswid,
           the.ownership = !!the.ownership,
           Date = my(paste(Month,"-",the_year)))
  return(df_custom)
}
```

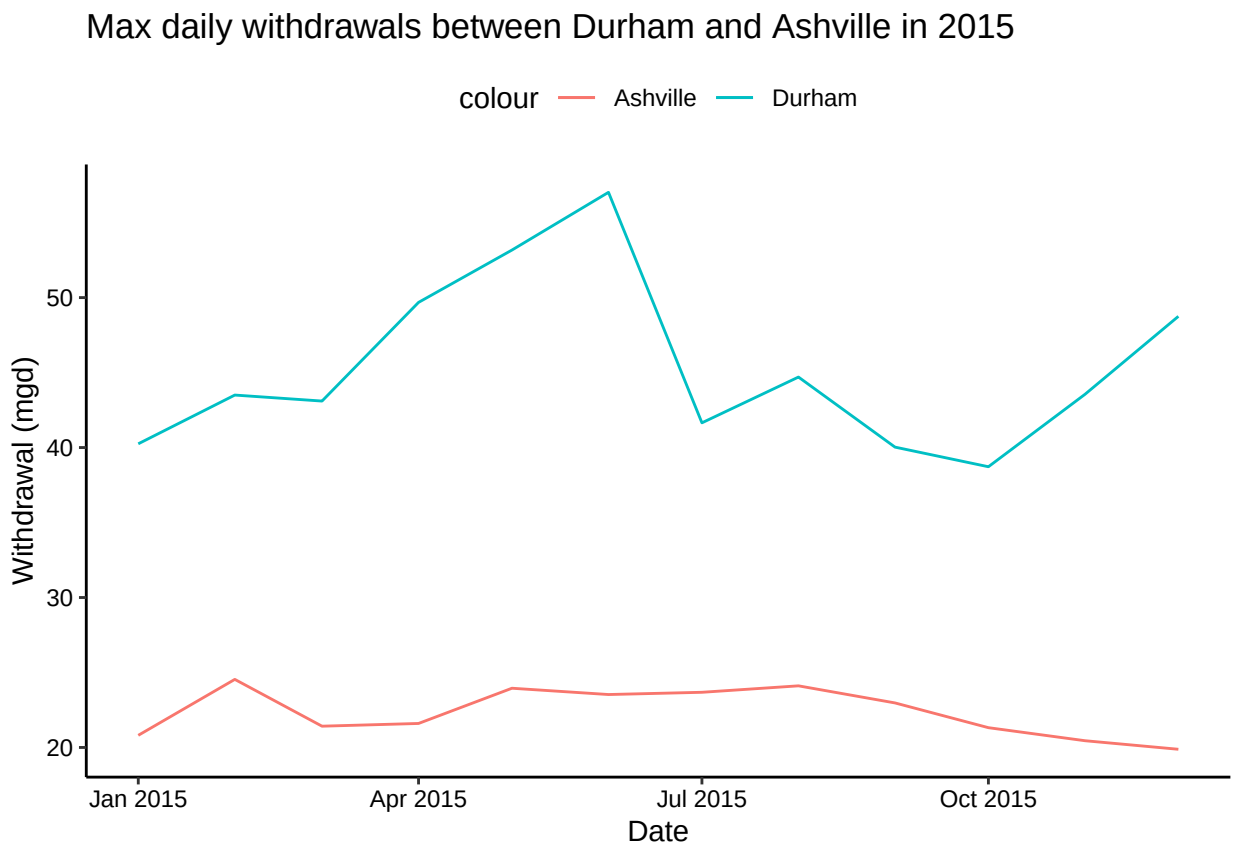
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID='03-32-010') for each month in 2015

```
#7
df_2015<-scrape.it(2015,'03-32-010')
view(df_2015)

ggplot(df_2015,aes(x= Date,y=max.withdrawals.mgd)) +
  geom_line() +
  labs(title = "Max daily withdrawals across the months for 2020",
       y="Withdrawal (mgd)",
       x="Date")
```



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.

```r
#8
df_Ash<-scrape.it(2015,'01-11-010')
view(df_Ash)

ggplot(df_Ash,aes(Date)) +
  geom_line(aes(y=max.withdrawals.mgd,color = "Ashville")) +
  geom_line(aes(y=df_2015$max.withdrawals.mgd, color = "Durham"))+
  labs(title = "Max daily withdrawals between Durham and Ashville in 2015",
       y="Withdrawal (mgd)",
       x="Date")
```

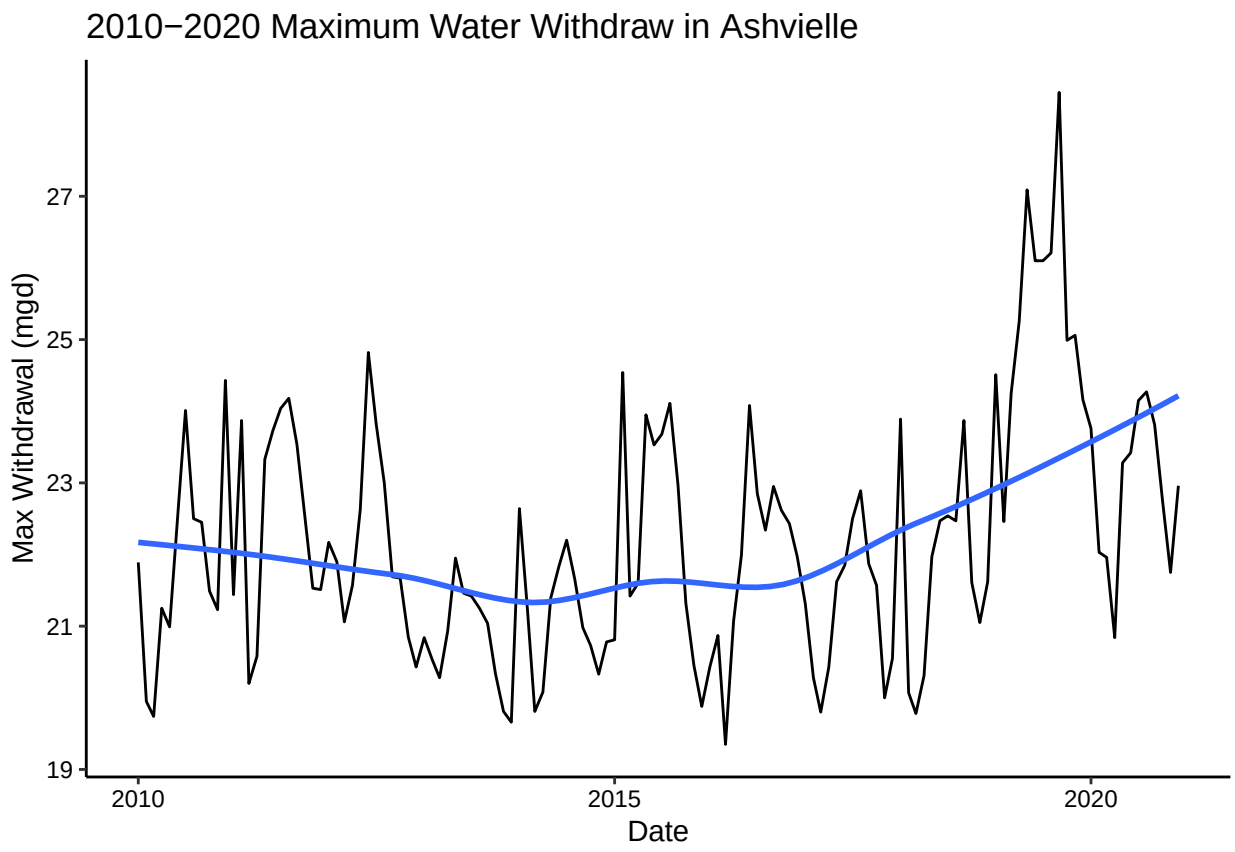


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2020.Add a smoothed line to the plot.

```r
#9
the_years = rep(2010:2020)
Ashville = '01-11-010'

the_dfs <- lapply(X = the_years,
                  FUN = scrape.it,
                  the_pwsid=Ashville)

the_dfs <- map(the_years,scrape.it, the_pwsid=Ashville)

the_df <- bind_rows(the_dfs)

ggplot(the_df,aes(x=Date,y=max.withdrawals.mgd)) +
```

```
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2010-2020 Maximum Water Withdraw in Ashvielle"),
       y="Max 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, the water usage appears to increase over time.