Assignment 10: Data Scraping

Joshua Frear

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_10_Data_Scraping.Rmd") prior to submission.

The completed exercise is due on Tuesday, April 6 at 11:59 pm.

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] "D:/Data_872/Environmental_Data_Analytics_2021/Assignments"
```

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

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 2019 Municipal Local Water Supply Plan (LWSP):

- Navigate to https://www.ncwater.org/WUDC/app/LWSP/search.php
- Change the date from 2020 to 2019 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=2019

Indicate this website as the as the URL to be scraped.

```
#2
webpage <- read_html("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2019")
```

3. The data we want to collect are listed below:

- From the "System Information" section:
- Water system name
- PSWID
- Ownership
- From the "Water Supply Sources" section:
- Maximum monthly withdrawals (MGD)

In the code chunk below scrape these values into the supplied variable names.

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

water_system_name
```

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

```
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()

maxmonthlywdrawals <- webpage %>%
  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.)

   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...

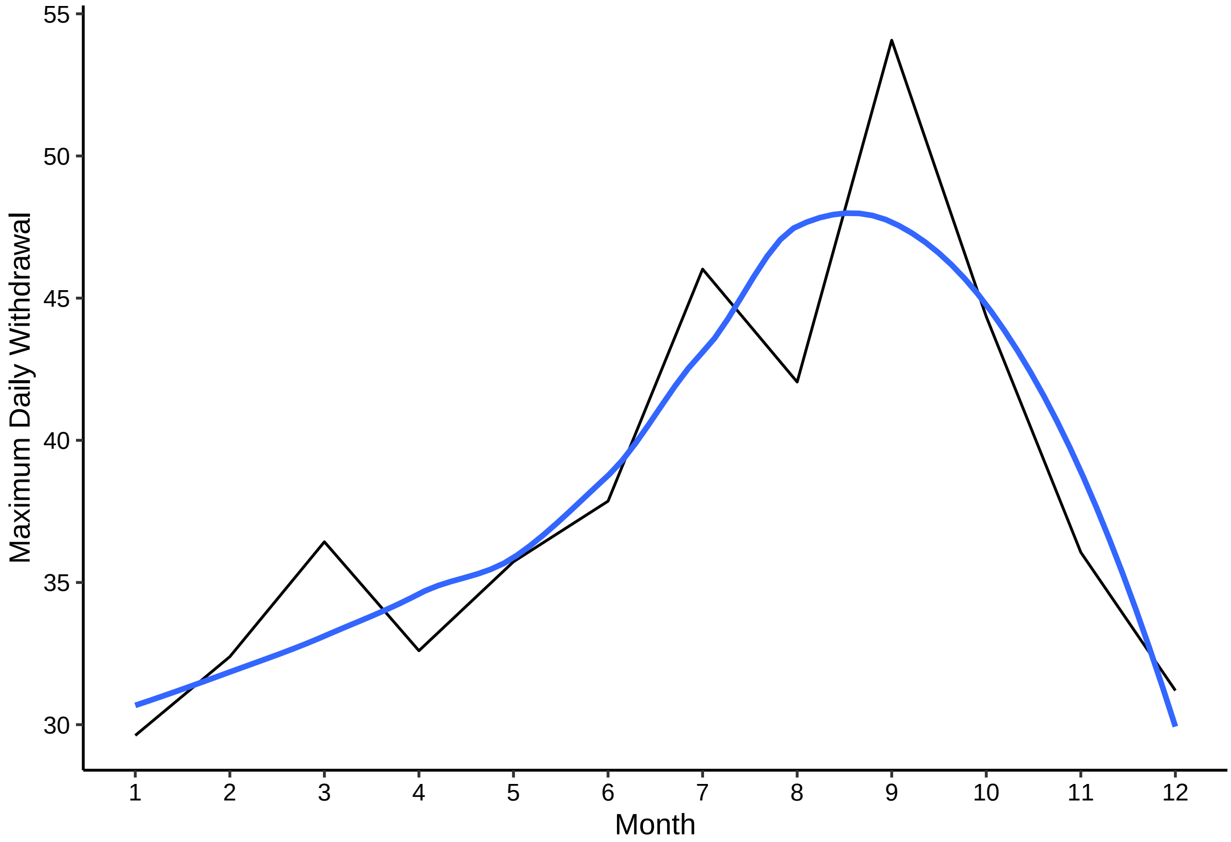
5. Plot the max daily withdrawals across the months for 2019.

```
#4
df_water <- data.frame("Month" = c(1,5,9,2,6,10,3,7,11,4,8,12),
                             "Year" = rep(2019,12),
                             "Max_Withdrawals_mgd" = as.numeric(maxmonthlywdrawals))
df_water <- arrange(df_water, Month)
df_water$Month <- month(df_water$Month)
```

```
df_water <- df_water %>%
  mutate(Water_system_name = !!water_system_name,
         PSWID = !!pswid,
         Ownership = !!ownership,
         Date = my(paste(Month,"/",Year)))

#5
plot1 <- ggplot(df_water,aes(x=Month,y=Max_Withdrawals_mgd)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) + scale_x_continuous(breaks=seq(1, 12)) +
  ylab("Maximum Daily Withdrawal")
print(plot1)
```

```
## `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 data scraped.

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

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

      #Set the element address variables (found with selector gadget)
```

```
    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)"
    maxmonthlywdrawals_tag <- "th~ td+ td"

    #Scrape
    the_water_system_name <- the_website %>% html_nodes(water_system_name_tag) %>% html_text()
    the_pswid <- the_website %>%   html_nodes(pswid_tag) %>%  html_text()
    the_ownership <- the_website %>% html_nodes(ownership_tag) %>% html_text()
    the_maxmonthlywdrawals <- the_website %>% html_nodes(maxmonthlywdrawals_tag) %>% html_text()

    #Convert to dataframe
    df_water <- data.frame("Month" = c(1,5,9,2,6,10,3,7,11,4,8,12),
                             "Year" = rep(the_year,12),
                             "Max_Withdrawals_mgd" = as.numeric(the_maxmonthlywdrawals))
    df_water <- arrange(df_water, Month)
    df_water$Month <- month(df_water$Month)

    df_water <- df_water %>%
      mutate(Water_system_name = !!the_water_system_name,
         PSWID = !!the_pswid,
         Ownership = !!the_ownership,
         Date = my(paste(Month,"/",Year))) %>%
      relocate(Date, .before = Max_Withdrawals_mgd)

  Sys.sleep(1)

  #Return dataframe
  return(df_water)
}
```

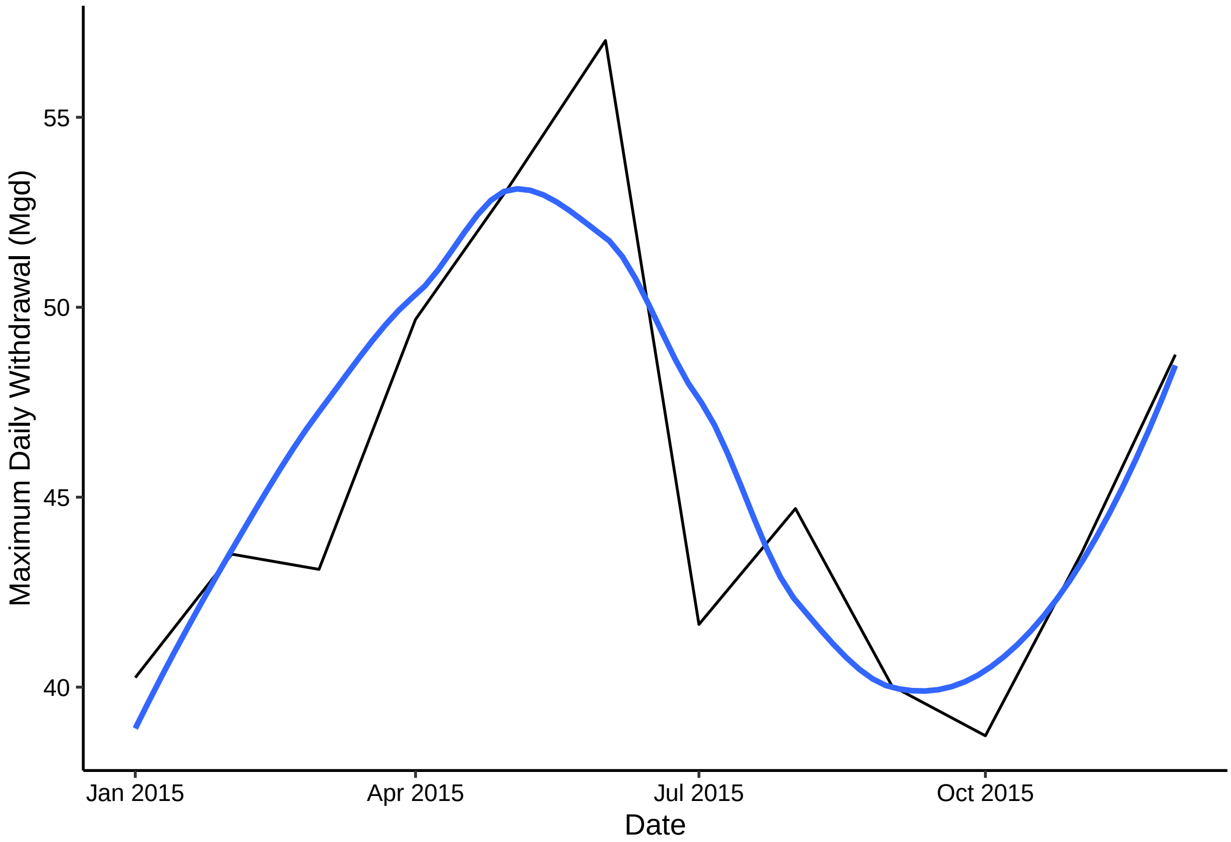
7. Use the function above to extract and plot max daily withdrawals for Durham for each month in 2015

```
#7
durham_2015 <- scrape.water(2015, "03-32-010")
durham_2015
```

```
##    Month Year       Date Max_Withdrawals_mgd Water_system_name      PSWID
## 1      1 2015 2015-01-01               40.25            Durham 03-32-010
## 2      2 2015 2015-02-01               43.50            Durham 03-32-010
## 3      3 2015 2015-03-01               43.10            Durham 03-32-010
## 4      4 2015 2015-04-01               49.68            Durham 03-32-010
## 5      5 2015 2015-05-01               53.17            Durham 03-32-010
## 6      6 2015 2015-06-01               57.02            Durham 03-32-010
## 7      7 2015 2015-07-01               41.65            Durham 03-32-010
## 8      8 2015 2015-08-01               44.70            Durham 03-32-010
## 9      9 2015 2015-09-01               40.03            Durham 03-32-010
## 10    10 2015 2015-10-01               38.72            Durham 03-32-010
## 11    11 2015 2015-11-01               43.55            Durham 03-32-010
## 12    12 2015 2015-12-01               48.75            Durham 03-32-010
##       Ownership
## 1  Municipality
## 2  Municipality
## 3  Municipality
## 4  Municipality
## 5  Municipality
## 6  Municipality
## 7  Municipality
## 8  Municipality
## 9  Municipality
## 10 Municipality
## 11 Municipality
## 12 Municipality
```

```
ggplot(durham_2015, aes(x = Date, y = Max_Withdrawals_mgd)) + geom_line() +
  geom_smooth(method = "loess", se = F) + ylab("Maximum Daily Withdrawal (Mgd)")
```

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



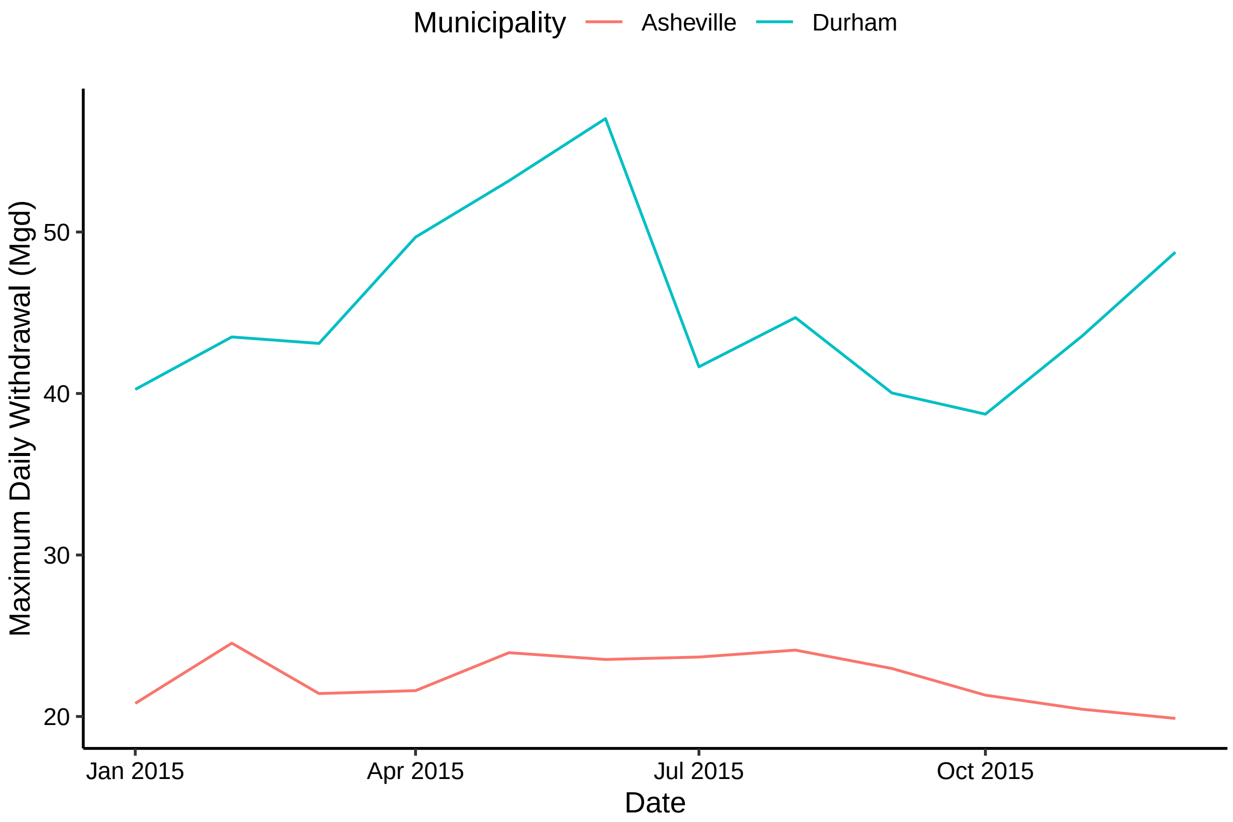
8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares the Asheville to Durham's water withdrawals.

```
#8
asheville_2015 <- scrape.water(2015, "01-11-010")
asheville_2015
```

```
##    Month Year       Date Max_Withdrawals_mgd Water_system_name     PSWID
## 1      1 2015 2015-01-01               20.81         Asheville 01-11-010
## 2      2 2015 2015-02-01               24.54         Asheville 01-11-010
## 3      3 2015 2015-03-01               21.42         Asheville 01-11-010
## 4      4 2015 2015-04-01               21.60         Asheville 01-11-010
## 5      5 2015 2015-05-01               23.95         Asheville 01-11-010
## 6      6 2015 2015-06-01               23.53         Asheville 01-11-010
## 7      7 2015 2015-07-01               23.68         Asheville 01-11-010
## 8      8 2015 2015-08-01               24.11         Asheville 01-11-010
## 9      9 2015 2015-09-01               22.97         Asheville 01-11-010
## 10    10 2015 2015-10-01               21.32         Asheville 01-11-010
## 11    11 2015 2015-11-01               20.45         Asheville 01-11-010
## 12    12 2015 2015-12-01               19.88         Asheville 01-11-010
##       Ownership
## 1  Municipality
## 2  Municipality
## 3  Municipality
## 4  Municipality
## 5  Municipality
## 6  Municipality
## 7  Municipality
## 8  Municipality
## 9  Municipality
## 10 Municipality
## 11 Municipality
## 12 Municipality
```

```
combined_df <- bind_rows(durham_2015, asheville_2015)

plot2 <- ggplot(combined_df, aes(x = Date, y = Max_Withdrawals_mgd)) +
  geom_line(aes(color = Water_system_name)) +
  ylab("Maximum Daily Withdrawal (Mgd)") + labs(color = "Municipality")
print(plot2)
```



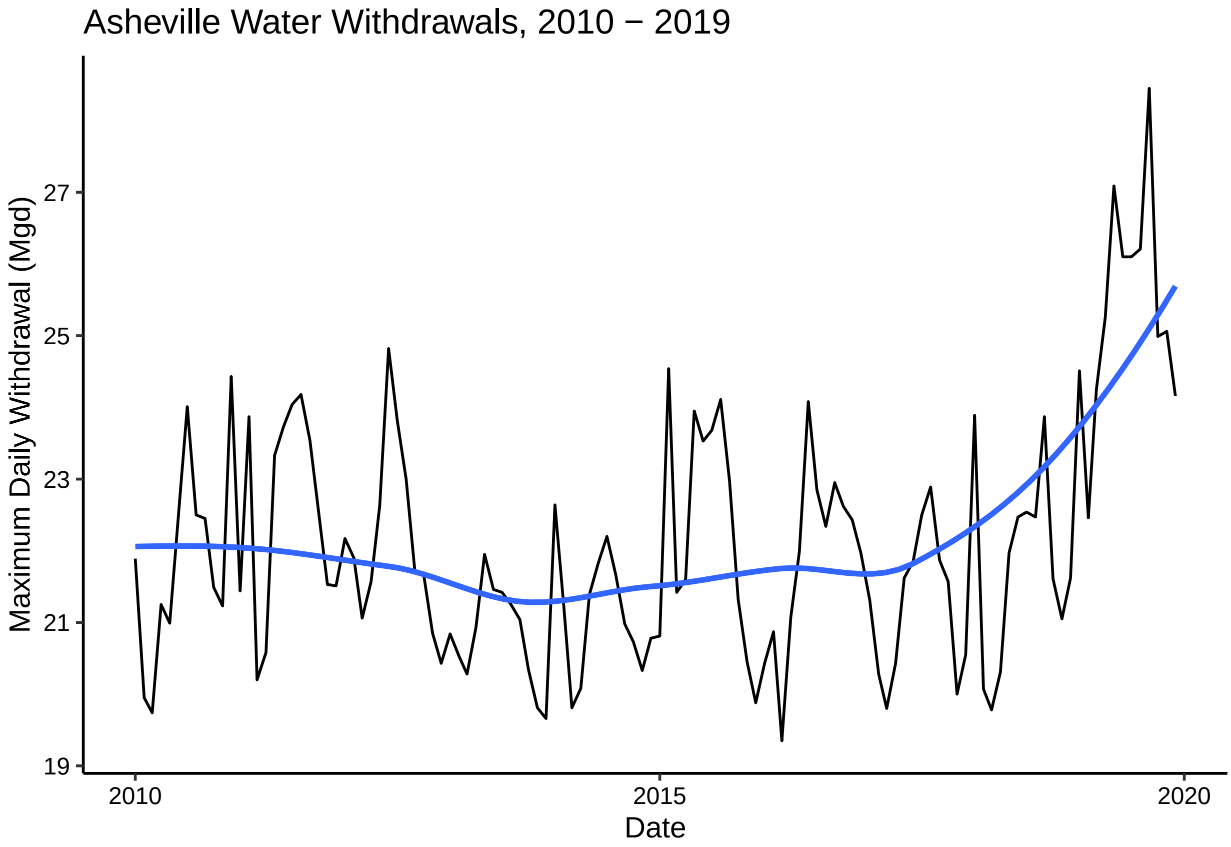
9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

```
#9
asheville_2010 <- scrape.water(2010, "01-11-010")
asheville_2011 <- scrape.water(2011, "01-11-010")
asheville_2012 <- scrape.water(2012, "01-11-010")
asheville_2013 <- scrape.water(2013, "01-11-010")
asheville_2014 <- scrape.water(2014, "01-11-010")
asheville_2015 <- scrape.water(2015, "01-11-010")
asheville_2016 <- scrape.water(2016, "01-11-010")
asheville_2017 <- scrape.water(2017, "01-11-010")
asheville_2018 <- scrape.water(2018, "01-11-010")
asheville_2019 <- scrape.water(2019, "01-11-010")

full_asheville <- bind_rows(asheville_2010,asheville_2011,asheville_2012,
                            asheville_2013,asheville_2014,asheville_2015,asheville_2016,
                            asheville_2017,asheville_2018,asheville_2019)

plot3 <- ggplot(full_asheville, aes(x = Date, y = Max_Withdrawals_mgd)) + geom_line() +
  geom_smooth(method = "loess", se = F) + ylab("Maximum Daily Withdrawal (Mgd)") +
  labs(title = "Asheville Water Withdrawals, 2010 - 2019")
print(plot3)
```

```
## `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, it looks like there has been a secular increasing trend from 2017 to 2019.