Assignment 10: Data Scraping

Anne Harshbarger

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] "C:/Users/Anne/Documents/ENV872/Environmental_Data_Analytics_2021/Assignments"
```

```
setwd("~/ENV872/Environmental_Data_Analytics_2021/Assignments")

#install.packages("tidyverse", "rvest", "ggplot2")
library(tidyverse)
library(rvest)
```

```
## Warning: package 'rvest' was built under R version 4.0.4
```

```
library(ggplot2)
library(lubridate)

mytheme <- theme_bw() + theme( #create ggplot theme
  text = element_text(color = "#000033", size = 12),
  axis.text = element_text(color = "black", size = 12),
  legend.position = "right",
  axis.line = element_line(color = "#000033", size = 1.25))
theme_set(mytheme) #set as default theme
```

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
theURL <- 'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2019'
webpage <- read_html(theURL) #set webpage to be scraped
```

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
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
monthly_withdrawals_max <- 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.

```r
#4
durham_lwsp <- data.frame(
    SystemName = system_name,
    PSWID = PSWID,
    Ownership = ownership,
    MaxWithdrawal= as.numeric(monthly_withdrawals_max), #add scraped data
    Month = as.character(c("Jan", "May", "Sep", "Feb", "Jun", "Oct",
                           "Mar", "Jul", "Nov", "Apr", "Aug", "Dec")), #add month
    Year = "2019" #add year
    )
durham_lwsp <-
  durham_lwsp %>%
  mutate(Date = my(paste0(Month, " ", Year))) #add date, put in date format (my, lubridate)

#5
#plot durham, 2019 data
monthly_wd_plot <- ggplot(data = durham_lwsp, aes(x=Date, y=MaxWithdrawal)) +
  geom_point() +
  geom_line() +
  scale_x_date(date_breaks = "1 month") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x = "Date (YYYY-MM-DD)", y = "Maximum Monthly Withdrawals (MGD)",
       title = "Maximum Monthly Withdrawals in Durham, 2019")
monthly_wd_plot
```

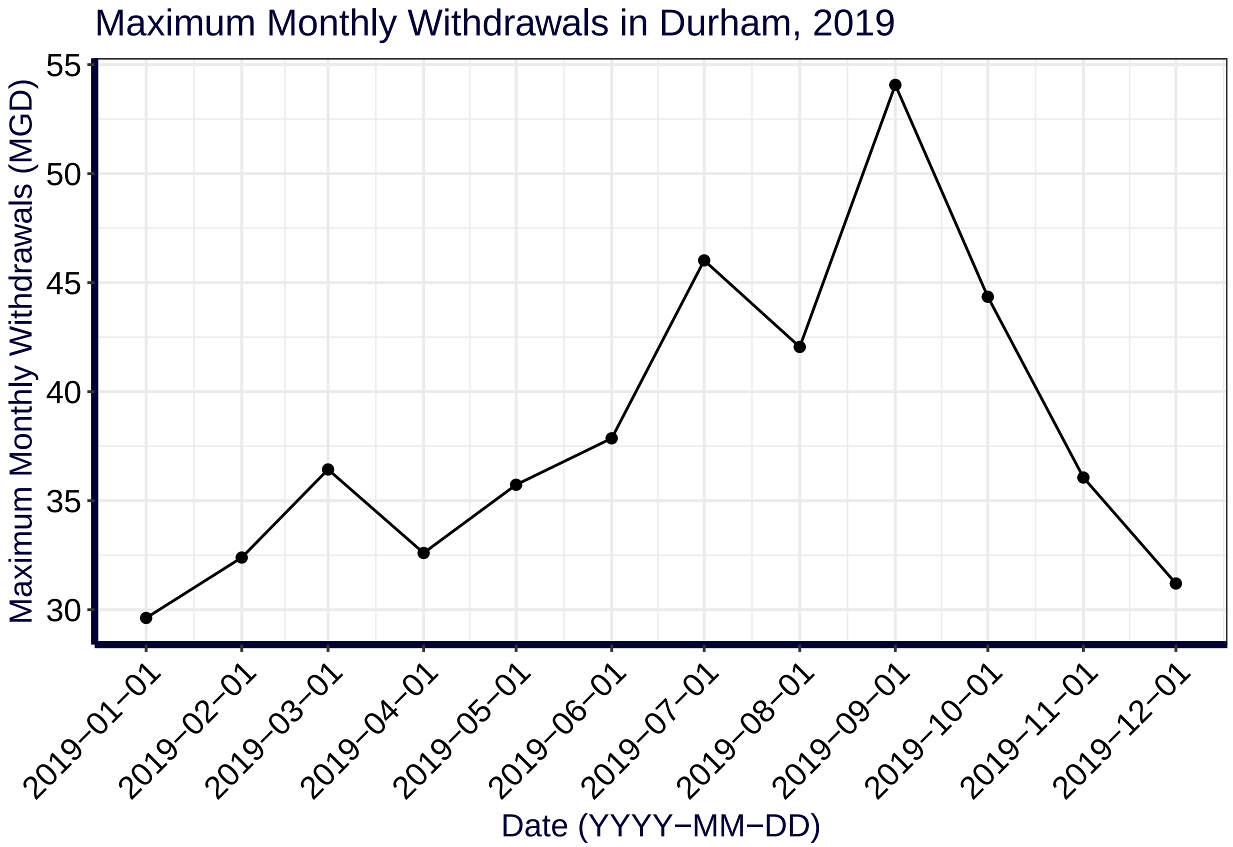

Figure 1: Maximum daily water withdrawals in Durham, NC for each month in 2019

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_deq_data <- function(the_PWSID, the_year){
  new_URL <- paste0("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=",
                   the_PWSID, "&year=", the_year) #url will populate with inputs
  new_webpage <- read_html(new_URL) #set webpage to scrape
  new_system_name <- new_webpage %>%
    html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>%
    html_text
  new_PSWID <- new_webpage %>%
    html_nodes('td tr:nth-child(1) td:nth-child(5)') %>%
    html_text
  new_ownership <- new_webpage %>%
    html_nodes('div+ table tr:nth-child(2) td:nth-child(4)') %>%
    html_text
  new_monthly_withdrawals_max <- new_webpage %>%
    html_nodes('th~ td+ td') %>%
    html_text #scrape variables
  new_lwsp <- data.frame(
    SystemName = new_system_name,
    PSWID = new_PSWID,
    Ownership = new_ownership,
    MaxWithdrawal= as.numeric(new_monthly_withdrawals_max), #add scraped data
    Month = as.character(c("Jan", "May", "Sep", "Feb", "Jun", "Oct",
                           "Mar", "Jul", "Nov", "Apr", "Aug", "Dec")), #add month
    Year = the_year #add year
    )
  new_lwsp <- new_lwsp %>%
    mutate(Date = my(paste0(Month, " ", Year))) #add date, put in date format (my)
  return(new_lwsp)

}
```

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

```
#7
#run custom function for durham, 2015
durham_lwsp_2015 <- scrape_deq_data(the_PWSID = "03-32-010", the_year = "2015")

#create plot with 2015 durham data
monthly_wd_2015_plot <- ggplot(data = durham_lwsp_2015, aes(x=Date, y=MaxWithdrawal)) +
  geom_point() +
  geom_line() +
  scale_x_date(date_breaks = "1 month") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x = "Date (YYYY-MM-DD)", y = "Maximum Monthly Withdrawals (MGD)",
       title = "Maximum Monthly Withdrawals in Durham, 2015")
monthly_wd_2015_plot
```

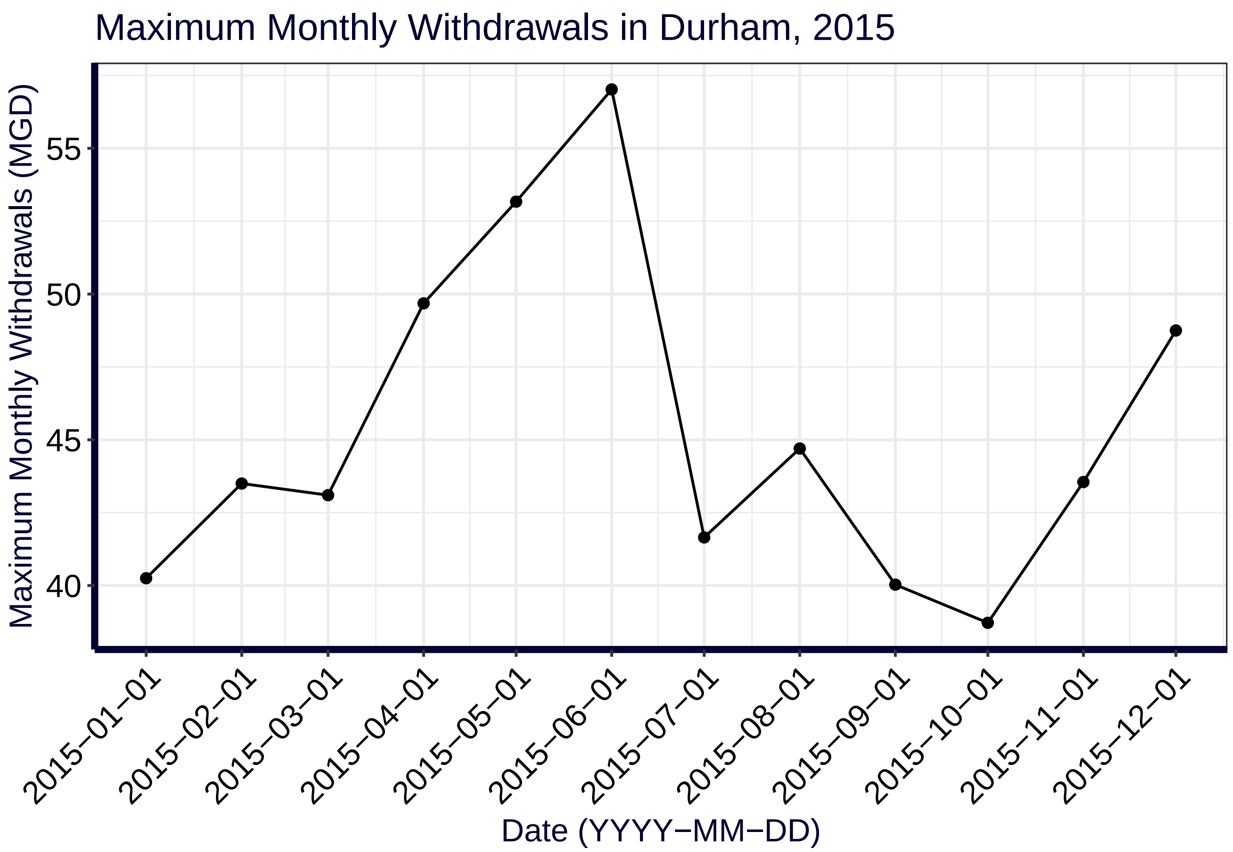

Figure 2: Maximum daily water withdrawals in Durham, NC for each month in 2015

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 the Asheville to Durham's water withdrawals.

```
#8
#run custom function to get asheville data
asheville_lwsp_2015 <- scrape_deq_data(the_PWSID = "01-11-010", the_year = "2015")

#bind durham 2015 data and asheville 2015 data
lwsp_2015_combined <- rbind(durham_lwsp_2015, asheville_lwsp_2015)

#create plot with one line for each city (color = SystemName)
monthly_wd_2015_combined_plot <- ggplot(data = lwsp_2015_combined,
                                        aes(x=Date, y=MaxWithdrawal, color = SystemName)) +
  geom_point() +
  geom_line() +
  scale_x_date(date_breaks = "1 month") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x = "Date (YYYY-MM-DD)",
       y = "Maximum Monthly Withdrawals (MGD)",
       title = "Maximum Monthly Withdrawals in Durham and Asheville, 2015",
       color = "System Name")
monthly_wd_2015_combined_plot
```

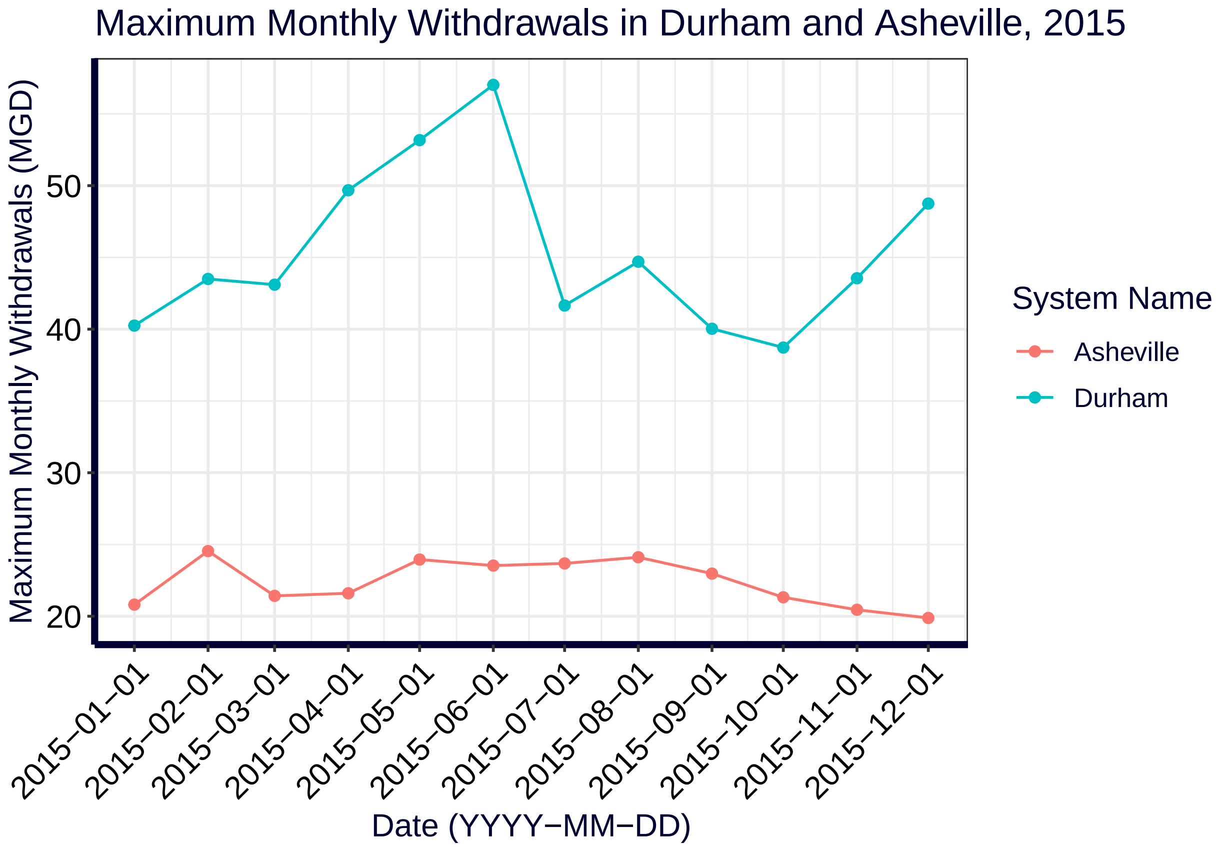

Figure 3: Maximum daily water withdrawals in Durham, NC (blue) and Asheville, NC (pink) for each month in 2015.

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
#run custom function on each year 2010-2019
asheville_lwsp_2010 <- scrape_deq_data(the_PWSID = "01-11-010", the_year = "2010")
asheville_lwsp_2011 <- scrape_deq_data(the_PWSID = "01-11-010", the_year = "2011")
asheville_lwsp_2012 <- scrape_deq_data(the_PWSID = "01-11-010", the_year = "2012")
asheville_lwsp_2013 <- scrape_deq_data(the_PWSID = "01-11-010", the_year = "2013")
asheville_lwsp_2014 <- scrape_deq_data(the_PWSID = "01-11-010", the_year = "2014")
asheville_lwsp_2015 <- scrape_deq_data(the_PWSID = "01-11-010", the_year = "2015")
asheville_lwsp_2016 <- scrape_deq_data(the_PWSID = "01-11-010", the_year = "2016")
asheville_lwsp_2017 <- scrape_deq_data(the_PWSID = "01-11-010", the_year = "2017")
asheville_lwsp_2018 <- scrape_deq_data(the_PWSID = "01-11-010", the_year = "2018")
asheville_lwsp_2019 <- scrape_deq_data(the_PWSID = "01-11-010", the_year = "2019")

#bind into one data frame
asheville_lwsp_decade <- rbind(asheville_lwsp_2010, asheville_lwsp_2011, asheville_lwsp_2012,
                               asheville_lwsp_2013, asheville_lwsp_2014, asheville_lwsp_2015,
                               asheville_lwsp_2016, asheville_lwsp_2017, asheville_lwsp_2018,
                               asheville_lwsp_2019)

#create plot
```

```
monthly_wd_asheville_plot <- ggplot(data = asheville_lwsp_decade,
                                        aes(x=Date, y=MaxWithdrawal)) +
  geom_point() +
  geom_line() +
  geom_smooth(method = lm) +
  scale_x_date(date_breaks = "6 months") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x = "Date (YYYY-MM-DD)", y = "Maximum Monthly Withdrawals (MGD)",
       title = "Maximum Monthly Withdrawals in Asheville, 2010-2019")
monthly_wd_asheville_plot
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```

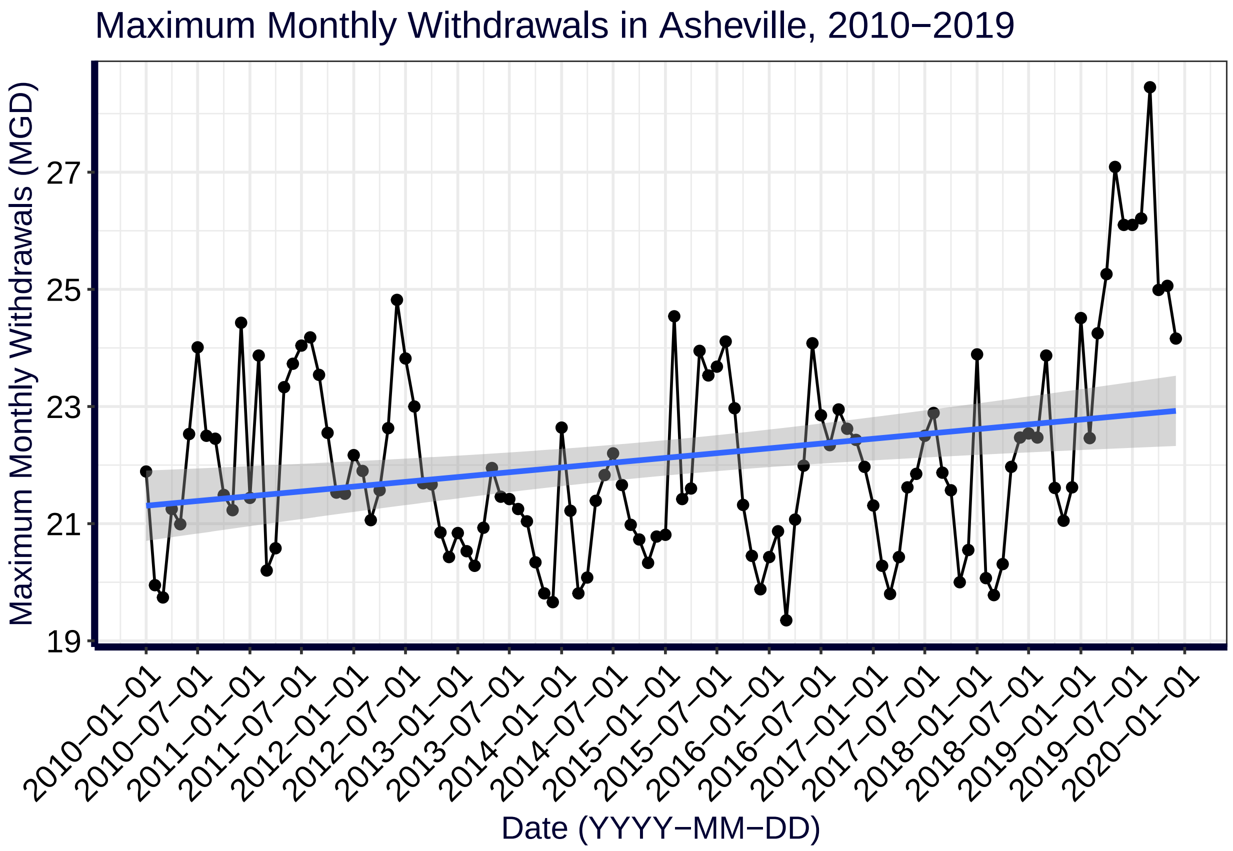

Figure 4: Maximum daily water withdrawals in Asheville, NC for each month between January 2010 and December 2019.

> Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time?

Yes, it appears that water usage in Asheville is increasing over time; however, the water usage was much higher in 2019 than the previous years, so this year could be an outlier. Further analysis is needed to determine if 2019 was unusually high or part of an increasing trend.

Duke Community Standard affirmation: I have adhered to the Duke Community Standard in completing this assignment. -Anne Harshbarger