Assignment 09: Data Scraping

Rachel Schoenecker

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/rscho/OneDrive/Documents/MEM Courses/Sem2/ENV 872/Environmental_Data_Analytics_2022/As
```

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

```
## Warning: package 'tidyverse' was built under R version 4.1.2
```

```
## Warning: package 'ggplot2' was built under R version 4.1.1
```

```
## Warning: package 'readr' was built under R version 4.1.1
```

```
## Warning: package 'purrr' was built under R version 4.1.1
```

```
## Warning: package 'forcats' was built under R version 4.1.1
```

```
## Warning: package 'rvest' was built under R version 4.1.3
```

```
## Warning: package 'lubridate' was built under R version 4.1.2
```

```
mytheme <- theme_bw(base_size = 10) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "right")
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=2020

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=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:
- Average 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 <- 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+ td") %>%
  html_text()
max.withdrawals.mgd <- as.numeric(max.withdrawals.mgd)
```

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.)

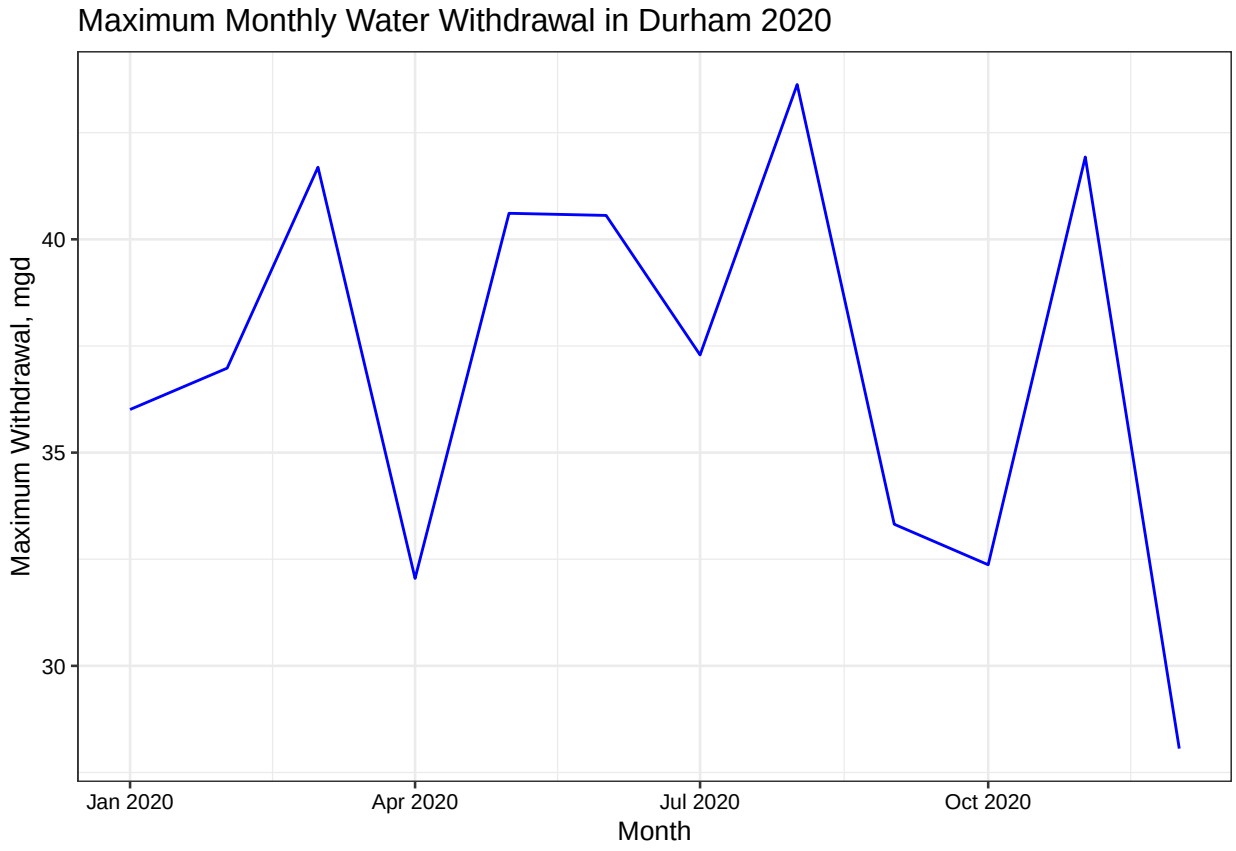
   > TIP: Use `rep()` to repeat a value when creating a dataframe.
   >
   > NOTE: It's likely you won't be able to scrape the monthly widthrawal data in order. You can overcome this by creating a month column in the same order the data are scraped: Jan, May, Sept, Feb, etc. . .

5. Plot the max daily withdrawals across the months for 2020

```r
#4
the_df <- data.frame(
  "water.system.name" = rep(water.system.name, 12),
  "pswid" = rep(pswid, 12),
  "ownership" = rep(ownership, 12),
  "monthly max withdrawal" = max.withdrawals.mgd,
  "month" = c(1, 5, 9, 2, 6, 10, 3, 7, 11,
              4, 8, 12),
  "year" = rep(2020, 12)
)

the_df <- the_df %>%
  mutate("date" = my(paste0(month,"/",year)))

#5
plot1 <- ggplot(the_df) + geom_line(aes(x = sort(date, decreasing = F),
                                        y = monthly.max.withdrawal), col = "blue") +
  labs(x = "Month", y = "Maximum Withdrawal, mgd",
       title = "Maximum Monthly Water Withdrawal in Durham 2020")
plot1
```

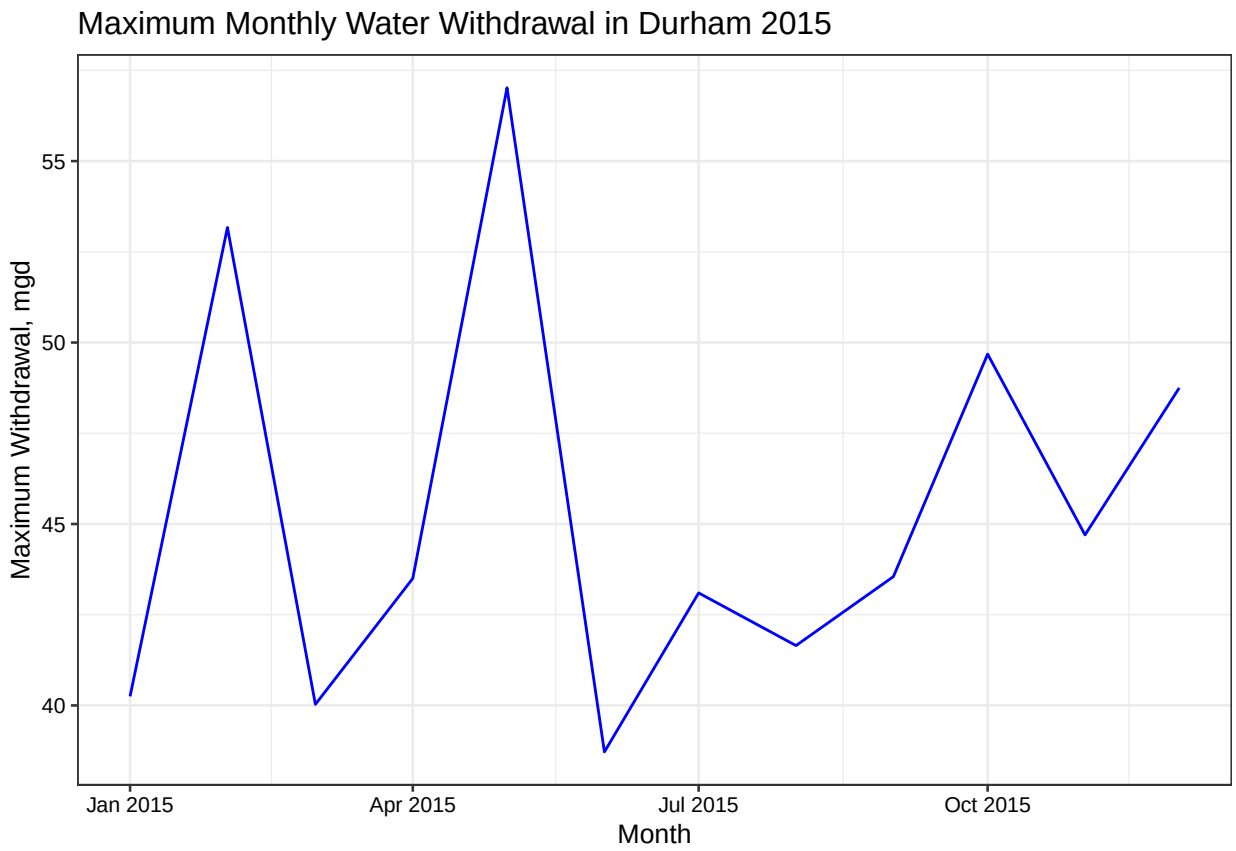


6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site scraped**.

```r
#6.
scrape.it <- function(the_year, the_pwsid){
  the_url <- ifelse(the_year==2020 & the_pwsid=="03-32-010",
                    'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2020', paste(
  webpage <- read_html(the_url)
  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+ td") %>%
  html_text()
  max.withdrawals.mgd <- as.numeric(max.withdrawals.mgd)
  the_df_2 <- data.frame(
  "water.system.name" = rep(water.system.name, 12),
  "pswid" = rep(pswid, 12),
  "ownership" = rep(ownership, 12),
  "monthly max withdrawal" = max.withdrawals.mgd,
  "month" = c(1, 5, 9, 2, 6, 10, 3, 7, 11,
              4, 8, 12),
  "year" = rep(the_year, 12))
  return(the_df_2)}
```

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

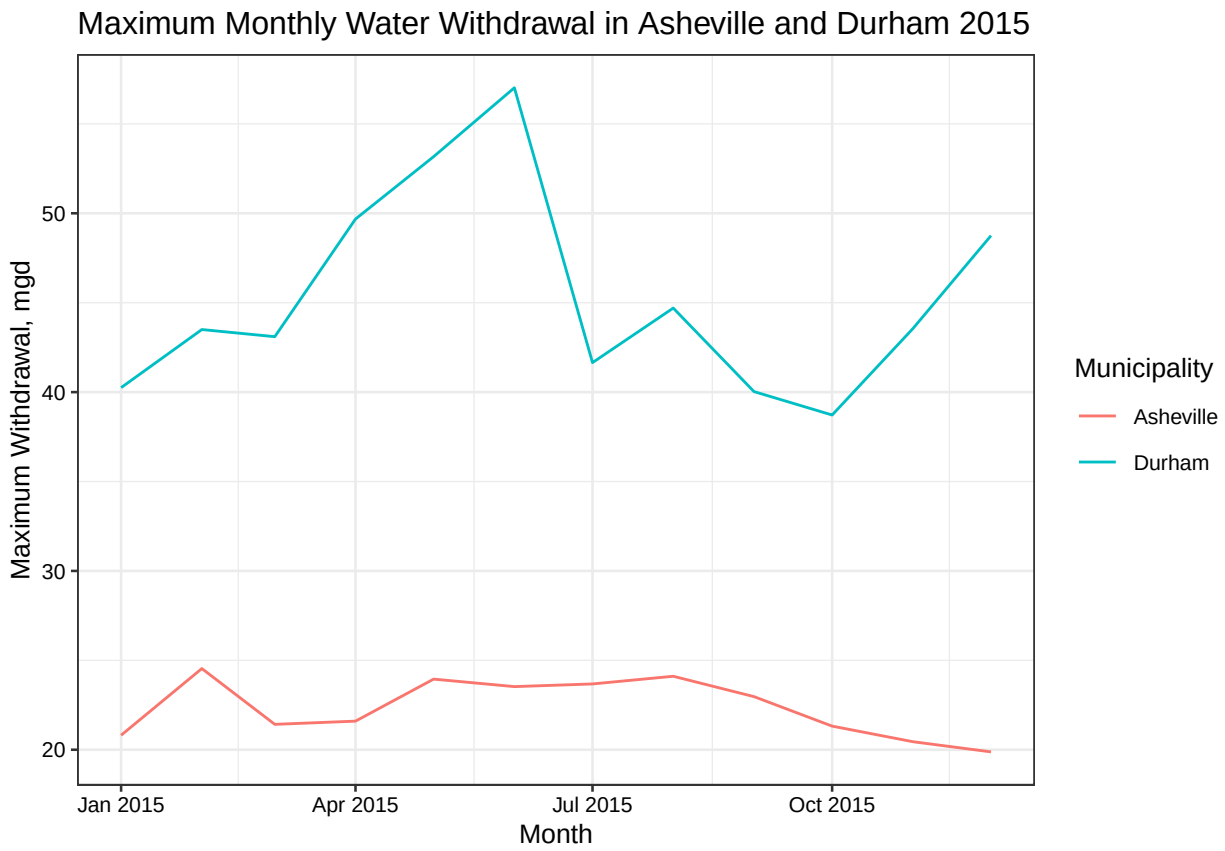
```r
#7
the_df2 <- scrape.it(the_year = 2015, the_pwsid = "03-32-010")
the_df2 <- the_df2 %>%
  mutate("date" = my(paste0(month,"/",year)))
plot2 <- ggplot(the_df2) + geom_line(aes(x = sort(date, decreasing = F),
                                y = monthly.max.withdrawal), col = "blue") +
  labs(x = "Month", y = "Maximum Withdrawal, mgd",
       title = "Maximum Monthly Water Withdrawal in Durham 2015")
plot2
```



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
the_df3 <- scrape.it(the_year = 2015, the_pwsid = "01-11-010")
the_df3 <- the_df3 %>%
  mutate("date" = my(paste0(month,"/",year)))
the_df_2015_comb <- rbind(the_df2, the_df3)

plot3 <- ggplot(the_df_2015_comb) + geom_line(aes(x = date,
                                                  y = monthly.max.withdrawal,
                                                  col = water.system.name)) +
  labs(x = "Month", y = "Maximum Withdrawal, mgd", col = "Municipality",
       title = "Maximum Monthly Water Withdrawal in Asheville and Durham 2015")
plot3
```

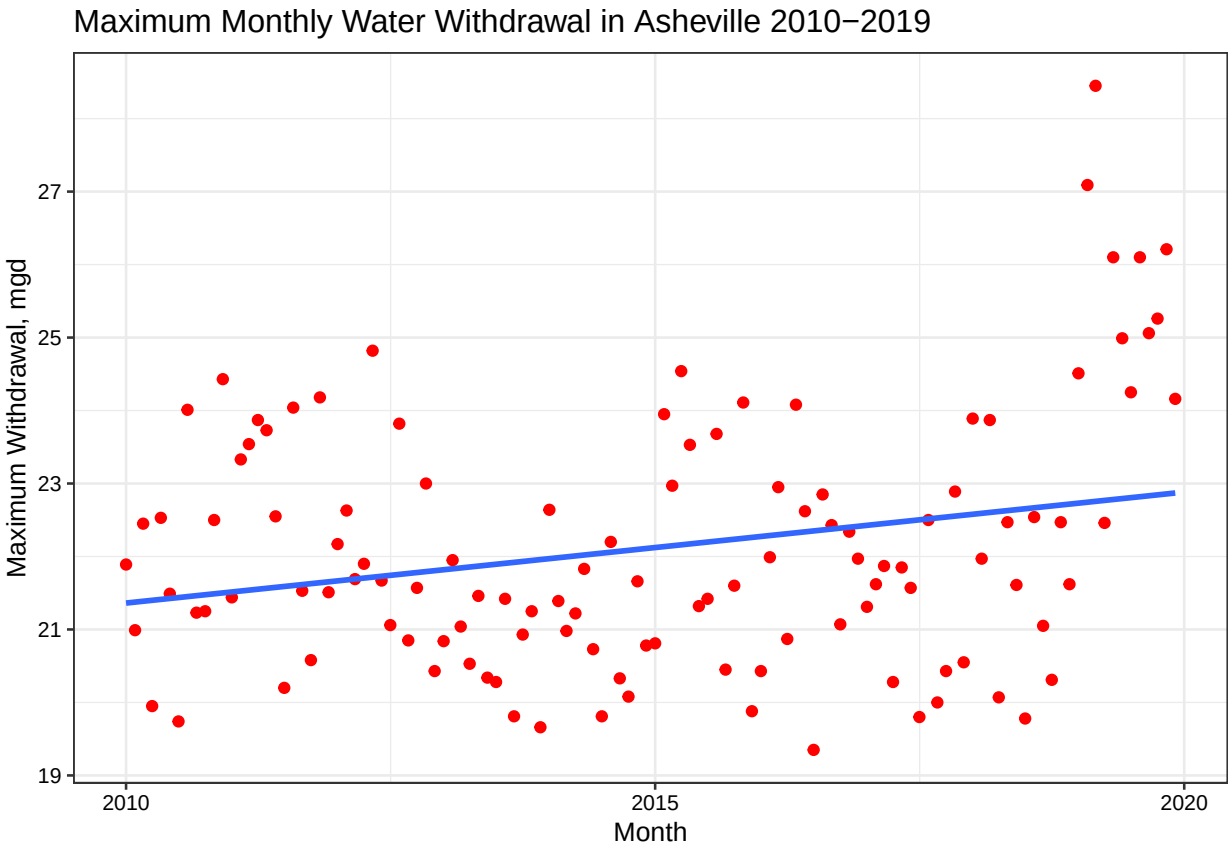


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

```
#9
years <- c(2010:2019)
the_df4 <- lapply(years, FUN = scrape.it, the_pwsid = "01-11-010") %>% bind_rows()
the_df4 <- the_df4 %>%
  mutate("date" = my(paste0(month,"/",year)))

plot4 <- ggplot(the_df4) + geom_point(aes(x = sort(date, decreasing = F),
                                y = monthly.max.withdrawal), col = "red") +
  geom_smooth(method = lm, aes(x = sort(date, decreasing = F),
                                y = monthly.max.withdrawal), se = F) +
  labs(x = "Month", y = "Maximum Withdrawal, mgd",
       title = "Maximum Monthly Water Withdrawal in Asheville 2010-2019")
plot4
```

```
## `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?
Answer: Yes, water usage in Asheville appears to have increased across the time period of interest.