Assignment 10: Data Scraping

Ana Bishop

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

getwd() #looks good
```

```
## [1] "/Users/anabishop/Documents/Data Analytics/ENV872"
```

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.)

```
#2
url <- 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 "36.1000".

```
#3
water.system.name <- html_nodes(url, 'div+ table tr:nth-child(1) td:nth-child(2)') %>% html_text()

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

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

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

Month <- html_nodes(url, '.fancy-table:nth-child(31) tr+ tr th') %>% html_text()
```

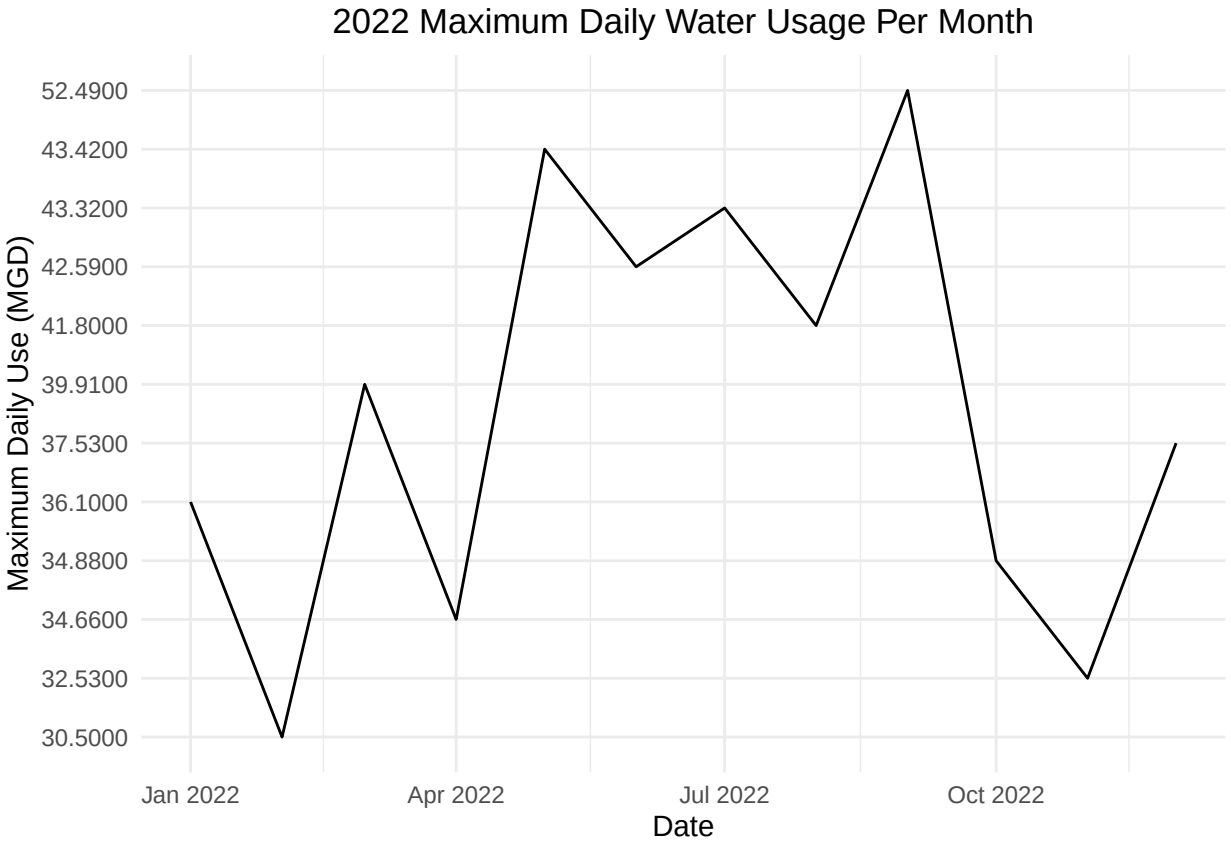
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   > 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 maximum daily withdrawals across the months for 2022

```r
#4
df_withdrawals <- data.frame("WaterSystemName" = rep(water.system.name),
                             "PWSID" = rep(PWSID),
                             "Ownership" = rep(ownership),
                              "Year" = rep(2022),
                              "Month" = (Month),
                             "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, group = 1)) +
  geom_line() +
  labs(title = "2022 Maximum Daily Water Usage Per Month", y = "Maximum Daily Use (MGD)") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



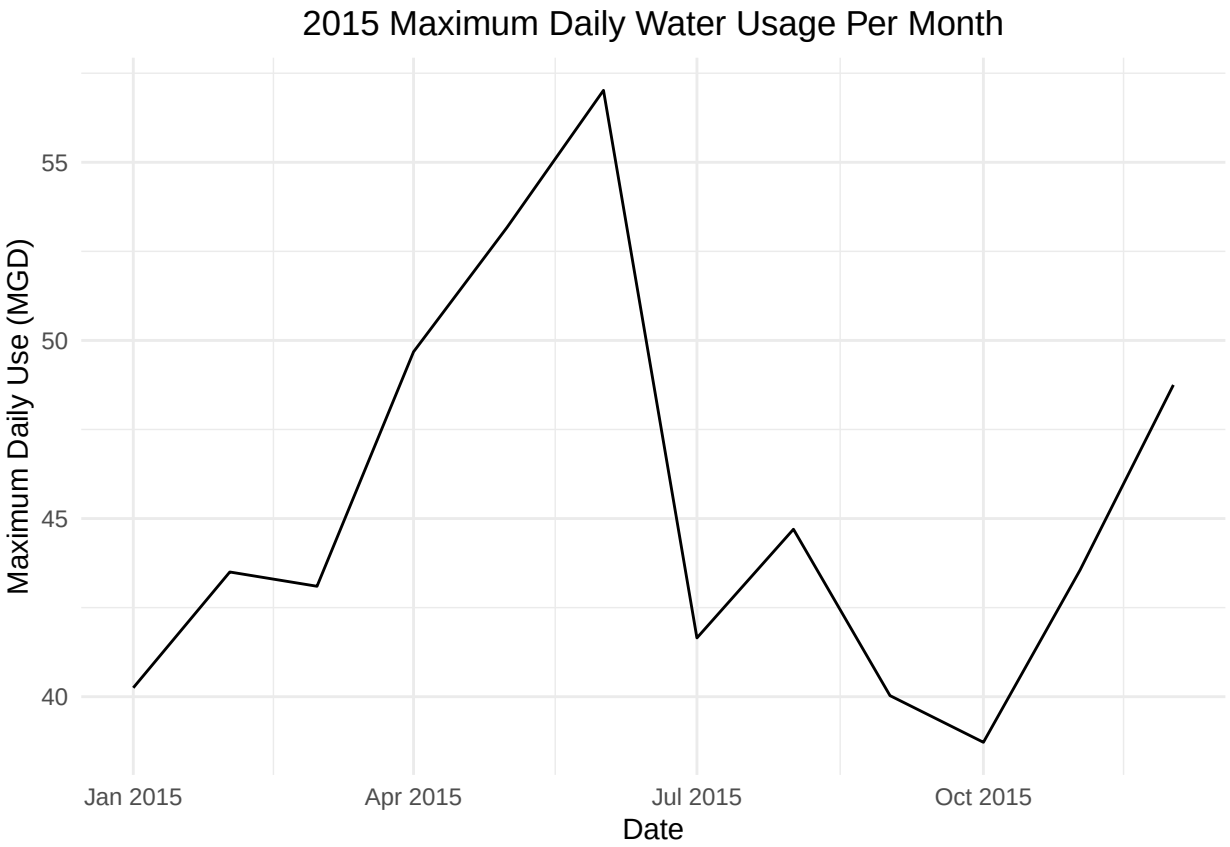
6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

```r
#6.
#Variables
year <- 2022
pwsid <- '03-32-010'
Month <- c('Jan', 'May', 'Sept', 'Feb', 'Jun', 'Oct', 'Mar', 'Jul', 'Nov', 'Apr', 'Aug', 'Dec')
```

```r
#Create scraping function
scrape.function <- function(year, pwsid){

  #Retrieve the website contents
  website <- read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?',
                              'pwsid=', pwsid, '&', 'year=', year))

  #Set the element address variables
  function_tag_water.system.name <- 'div+ table tr:nth-child(1) td:nth-child(2)'
  function_tag_PWSID <- 'td tr:nth-child(1) td:nth-child(5)'
  function_tag_ownership <- 'div+ table tr:nth-child(2) td:nth-child(4)'
  function_tag_max.withdrawals.mgd <- 'th~ td+ td'
  function_tag_Month <- '.fancy-table:nth-child(31) tr+ tr th'

  #Scrape the data items
  function_water.system.name <- website %>% html_nodes(function_tag_water.system.name) %>% html_text()
  function_PWSID <- website %>%   html_nodes(function_tag_PWSID) %>%  html_text()
  function_ownership <- website %>% html_nodes(function_tag_ownership) %>% html_text()
  function_max.withdrawals.mgd <- website %>% html_nodes(function_tag_max.withdrawals.mgd) %>% html_tex
  function_Month <- website %>% html_nodes(function_tag_Month) %>% html_text()

  #Convert to a dataframe
  df_withdrawals_function <- data.frame("Year" = rep(year,12),
                                "Month" = (Month),
                               "Max_Withdrawals_mgd" = as.numeric(function_max.withdrawals.mgd)) %>%
    mutate(Water_System_Name = !!function_water.system.name,
           PWSID = !!function_PWSID,
           Ownership = !!function_ownership,
           Date = my(paste(Month,"-",Year)))

  #Return the dataframe
  return(df_withdrawals_function)
}
```

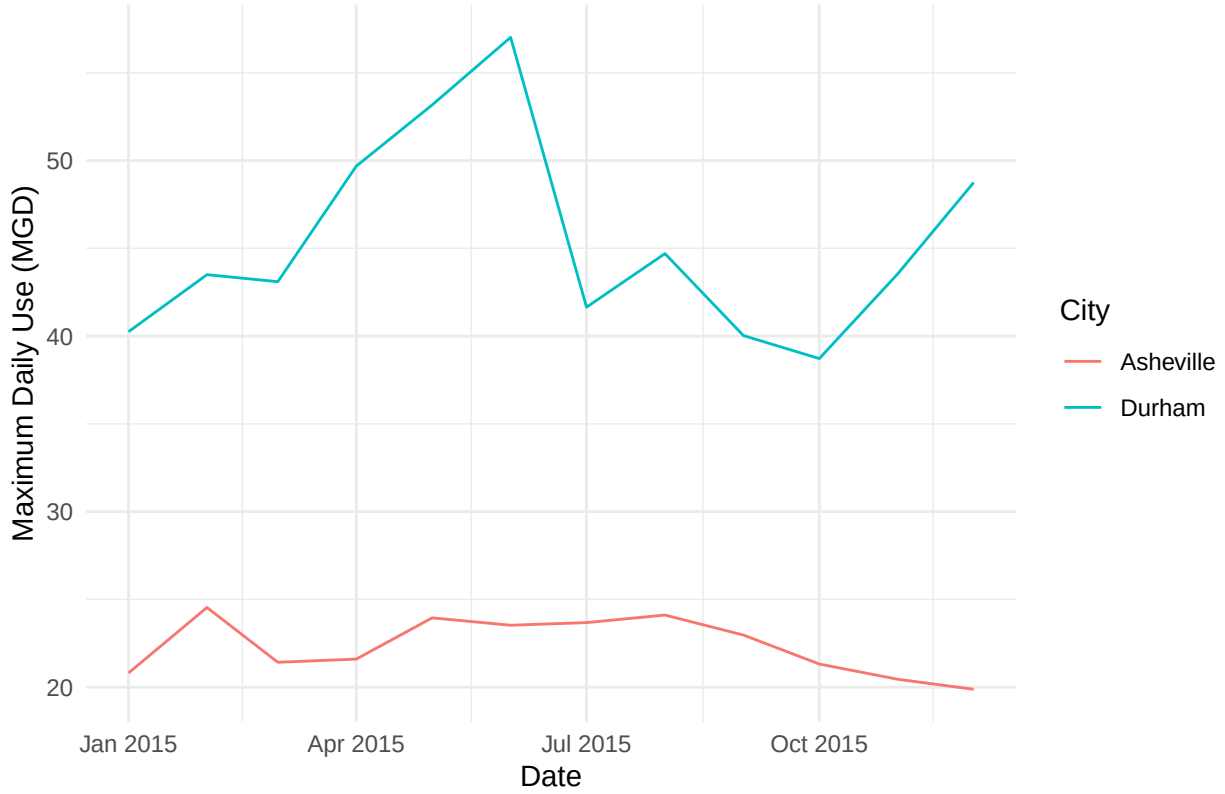
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID=‘03-32-010’) for each month in 2015

```r
#7
withdrawals_2015 <- scrape.function(2015, '03-32-010')

ggplot(withdrawals_2015, aes(x = Date, y = Max_Withdrawals_mgd, group = 1)) +
  geom_line() +
  labs(title = "2015 Maximum Daily Water Usage Per Month", y = "Maximum Daily Use (MGD)") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



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.

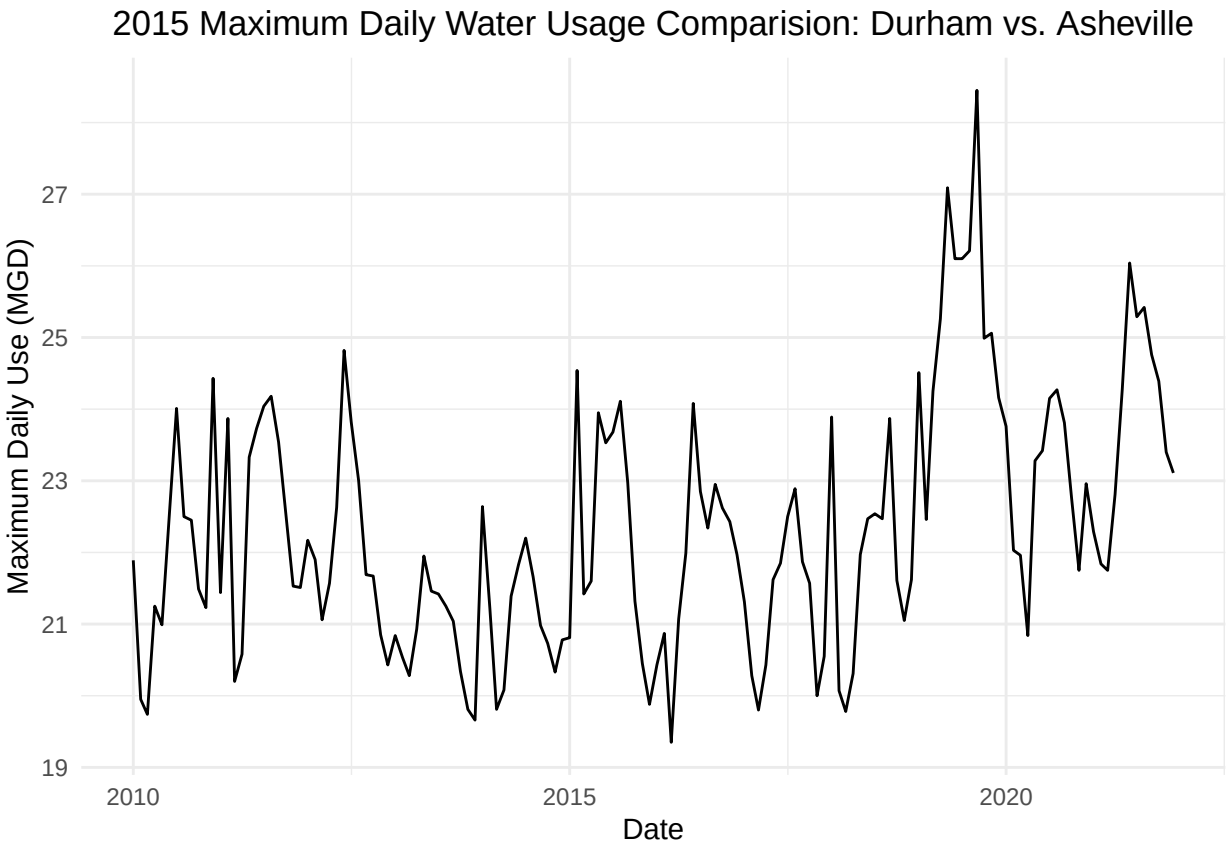
```
#8
withdrawals_asheville <- scrape.function(2015, '01-11-010')

comparison_2015 <- rbind(withdrawals_2015, withdrawals_asheville)

ggplot() +
  geom_line(data = comparison_2015, aes(x = Date, y = Max_Withdrawals_mgd, color = Water_System_Name)) +
  labs(title = "2015 Maximum Daily Water Usage Comparison Between Durham and Asheville", y = "Maximum Da
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



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
timeseries_years <-  c(2010:2021)

#Create a list of the year we want, the same length as the vector above
timeseries_pwsid <- rep.int("01-11-010",length(timeseries_years))

#"Map" the "scrape.it" function to retrieve data for all these
asheville_timeseries <- map2(timeseries_years, timeseries_pwsid, scrape.function) %>% bind_rows()

#Plot
ggplot() +
  geom_line(data = asheville_timeseries, aes(x = Date, y = Max_Withdrawals_mgd, group = 1)) +
  labs(title = "2015 Maximum Daily Water Usage Comparision: Durham vs. Asheville", y = "Maximum Daily Us
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



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 as if Asheville has an increasing trend, meaning that they are beginning to consistently increase their maximum water usage over the years.