Assignment 10: Data Scraping

Emma Childs

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(lubridate);library(viridis);library(here)

install.packages("rvest")
library(rvest)

install.packages("dataRetrieval")
library(dataRetrieval)
#specifically for NWIS data

install.packages("tidycensus")
library(tidycensus)

knitr::opts_chunk$set(tidy.opts = list(width.cutoff=80), tidy=TRUE)
```

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
Durham.webpage <- 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 "27.6400".

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

PWSID <- Durham.webpage %>%
    html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
    html_text()

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

max.withdrawals.mgd <- Durham.webpage %>%
    html_nodes("th~ td+ td") %>%
    html_text()
# the question asks for Max Day Use, but there isn't a value in the MGD columns
# that is equal to 27.6400 as described in the hint above. I highlighted the
# column that started with 36.1000 because that corresponds to the MDG I'm
# seeing on the water website.
```

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 average daily withdrawals across the months for 2022

```
# 4
max.withdrawals.month <- Durham.webpage %>%
    html_nodes("correct_code") %>%
    html_text()
# I think this is the right code, but I can't get the scraper to select the
# correct columns, grrr, so that specific scraper text is missing!

# max.daily.withdrawals.2022 <- 2022 Date <-
# as.Date(my(paste(max.withdrawals.month, '-', max.daily.withdrawals.2022)))
# class (Date) I can't get this chunk to run either because I can't get #4 to
# run without that correct scraper code.

# Durham.water.df <- data_frame('Water System Name' = water.system.name,
# 'Ownership' = ownership, 'Max Withdrawals Total' =
# as.numeric(max.withdrawals.mgd), 'Max Withdrawals Month' =
# max.withdrawals.month, 'Max Withdrawals Year' = max.withdrawals.year, 'Date'
# = as.Date(my(paste(max.withdrawals.month, '-', #max.withdrawals.year))) I
# can't run this because I still can't get my max.withdrawals.month to run, but
# I think I still understand the concept!

# 5 ggplot(Durham.water.df,aes(x=max.withdrawals.month,y=max.withdrawals.mgd))
# + geom_point() + labs(title = paste('2022 Max Withdrawals for Durham',
# water.system.name), subtitle = paste('PSWID', pswid), y='Max Withdrawals',
# x='Month')
```

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**.

```
# 6.
```

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

```
# 7
```

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
```

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
```

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