Assignment 09: Data Scraping

Ariel O'Callaghan

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on data scraping.

Directions

1. Rename this file `<FirstLast>_A09_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 to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.

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] "/home/guest/R/EDA-Fall2022/Assignments"
```

```
library(tidyverse)
library(lubridate)

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

# Set theme
mytheme <- theme_classic() +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
theme_set(mytheme)

knitr::opts_chunk$set(tidy.opts=list(width.cutoff=50), tidy=TRUE)
```

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham's 2021 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=2021

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=2021")
webpage
```

```
## {html_document}
## <html xmlns="http://www.w3.org/1999/xhtml" lang="en" xml:lang="en">
## [1] <head>\n<title>DWR :: Local Water Supply Planning</title>\n<meta http-equ ...
## [2] <body id="plan">\r\n<!--<div id="division-header">\r\n<a name="top" href= ...
```

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

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

```
pwsid <- webpage %>%
    html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
    html_text()
pwsid
```

```
## [1] "03-32-010"
```

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

```
## [1] "Municipality"
```

```
max.withdrawals.mgd <- webpage %>%
    html_nodes("th~ td+ td , th~ td+ td") %>%
    html_text()
max.withdrawals.mgd
```

```
##  [1] "27.6400" "41.7900" "36.7200" "27.9700" "37.9500" "42.2400" "30.5400"
##  [8] "43.6200" "31.2800" "33.7600" "46.0800" "29.7800"
```

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...

5. Create a line plot of the maximum daily withdrawals across the months for 2021

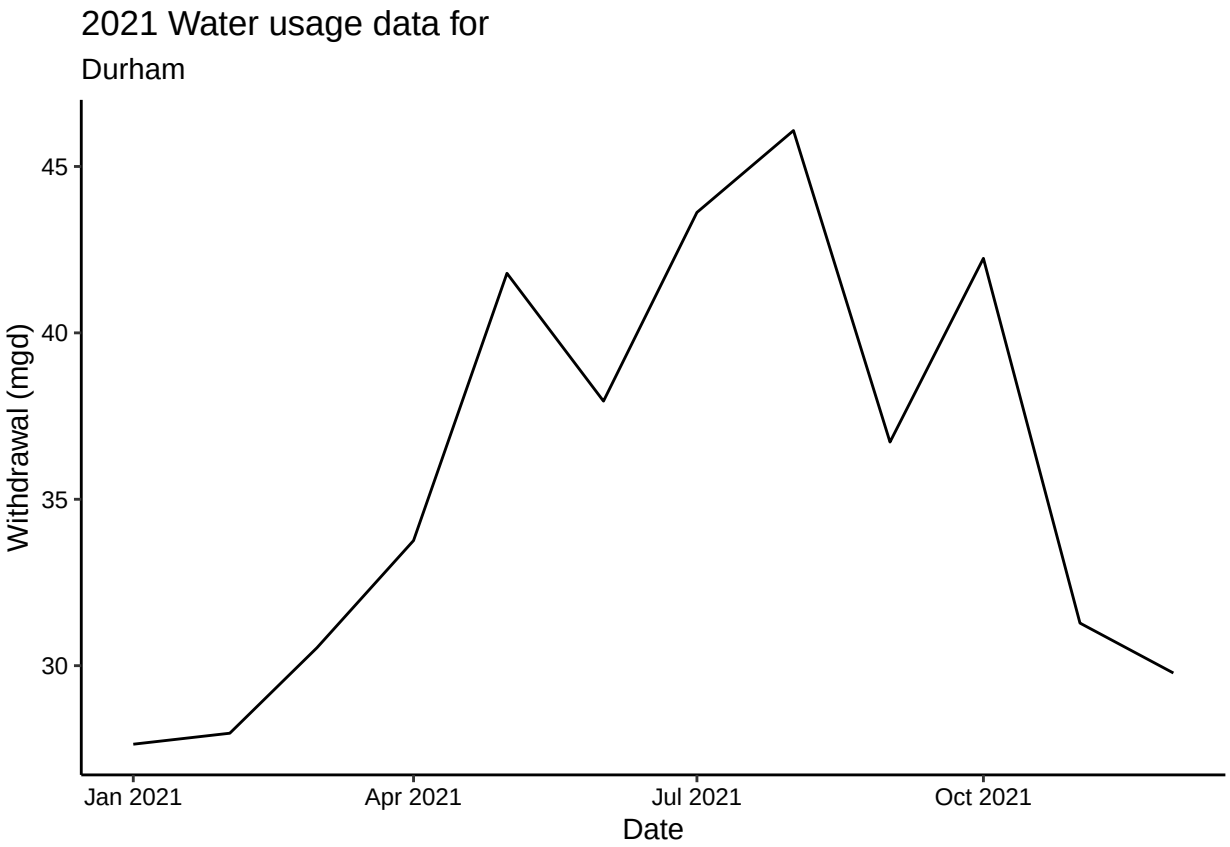
```
# 4 Order of months from scraped data. Month <-
# c(1,5,9,2,6,10,3,7,11,4,8,12)

df_withdrawals <- data.frame(Month = c(1, 5, 9, 2,
    6, 10, 3, 7, 11, 4, 8, 12), Year = rep(2021, 12),
    Max_Withdrawls_mgd = as.numeric(max.withdrawals.mgd))

df_withdrawals <- df_withdrawals %>%
    mutate(Water_system_name = !!water.system.name,
        PWSID = !!pwsid, Ownership = !!ownership, Date = my(paste(Month,
            "-", Year)))
class(Date)
```

```
## [1] "function"
```

```
# 5 plot geom_line(aes(group=1))
Durham.max.withdrawl.plot <- ggplot(df_withdrawals) +
    geom_line(aes(x = Date, y = as.numeric(max.withdrawals.mgd))) +
    labs(title = "2021 Water usage data for", subtitle = water.system.name,
        y = "Withdrawal (mgd)", x = "Date")
print(Durham.max.withdrawl.plot)
```



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.

scrape.it <- function(the_year, the_pwsid) {
    # https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2021'

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

    water.system.name <- the_website %>%
        html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
        html_text()
    water.system.name

    pwsid <- the_website %>%
        html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
        html_text()
    pwsid

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

```
    ownership

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

    # rep(1:12)
    df_scrape <- data.frame(Month = c(1, 5, 9, 2, 6,
        10, 3, 7, 11, 4, 8, 12), Year = rep(the_year,
        12), Max_Withdrawls_mgd = as.numeric(max.withdrawals.mgd)) %>%
        mutate(Water_system_name = !!water.system.name,
            PWSID = !!pwsid, Ownership = !!ownership,
            Date = my(paste(Month, "-", Year)))

    return(df_scrape)

}
```

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

```
# 7

df_durham_2015 <- scrape.it(2015, "03-32-010")
view(df_durham_2015)

df_durham_2015$Date <- as.Date(df_durham_2015$Date,
    format = "%Y-%M-%D")
class(df_durham_2015$Date)
```

```
## [1] "Date"
```

```
class(df_durham_2015$Max_Withdrawls_mgd)
```

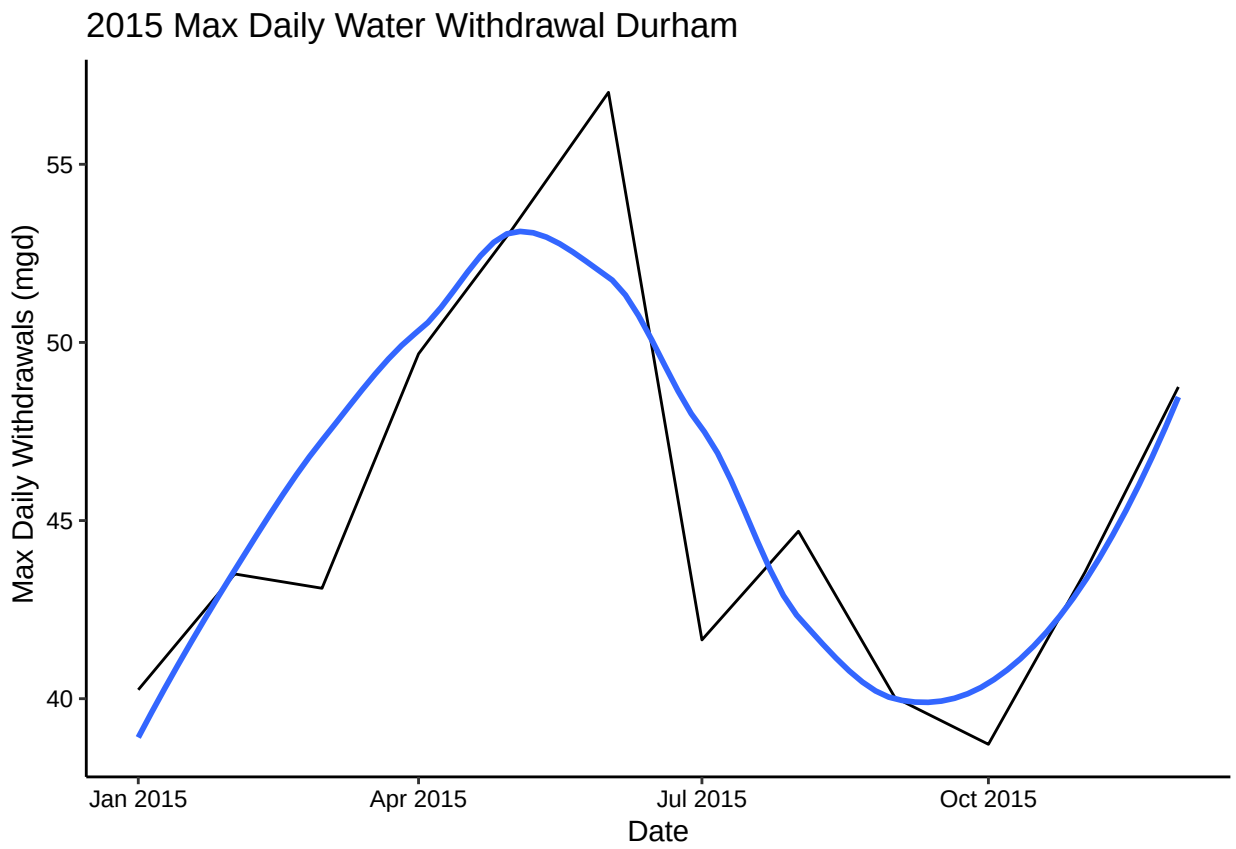
```
## [1] "numeric"
```

```
class(df_durham_2015$Date)
```

```
## [1] "Date"
```

```
Durham_2015_plot <- ggplot(df_durham_2015, aes(x = Date,
    y = Max_Withdrawls_mgd, group = 1)) + geom_line() +
    geom_smooth(method = "loess", se = FALSE) + labs(title = paste("2015 Max Daily Water Withdrawal Dur
    X = "Date", y = " Max Daily Withdrawals (mgd)")
print(Durham_2015_plot)
```

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




```
# geom_smooth(method='loess',se=FALSE)
```

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

df_Asheville_2015 <- scrape.it(2015, "01-11-010")
view(df_Asheville_2015)

durham_asheville <- rbind(df_Asheville_2015, df_durham_2015)


durham_asheville$Date <- as.Date(durham_asheville$Date,
    format = "%Y-%M-%D")
class(durham_asheville$Date)
```

```
## [1] "Date"
```

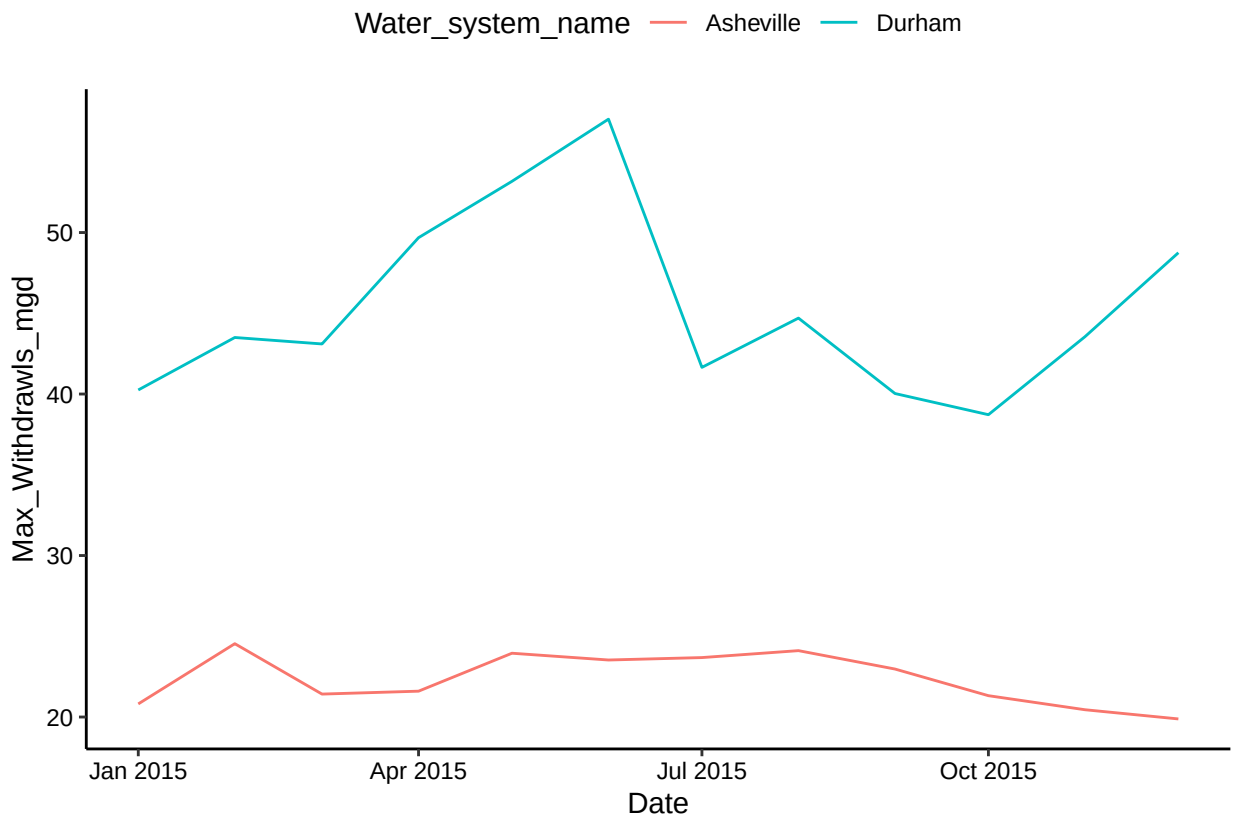
```
class(durham_asheville$Max_Withdrawls_mgd)
```

```
## [1] "numeric"
```

```
Durham__ashville_plot <- ggplot(durham_asheville, ) +
    geom_line(aes(x = Date, y = Max_Withdrawls_mgd,
        color = Water_system_name))
labs(title = paste("2015 Max Daily Water Withdrawal Durham and Asheville"),
    X = "Date", y = " Max Daily Withdrawals (mgd)")
```

```
## $X
## [1] "Date"
##
## $y
## [1] " Max Daily Withdrawals (mgd)"
##
## $title
## [1] "2015 Max Daily Water Withdrawal Durham and Asheville"
##
## attr(,"class")
## [1] "labels"
```

```
print(Durham__ashville_plot)
```



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.

   TIP: See Section 3.2 in the "09_Data_Scraping.Rmd" where we apply "map2()" to iteratively

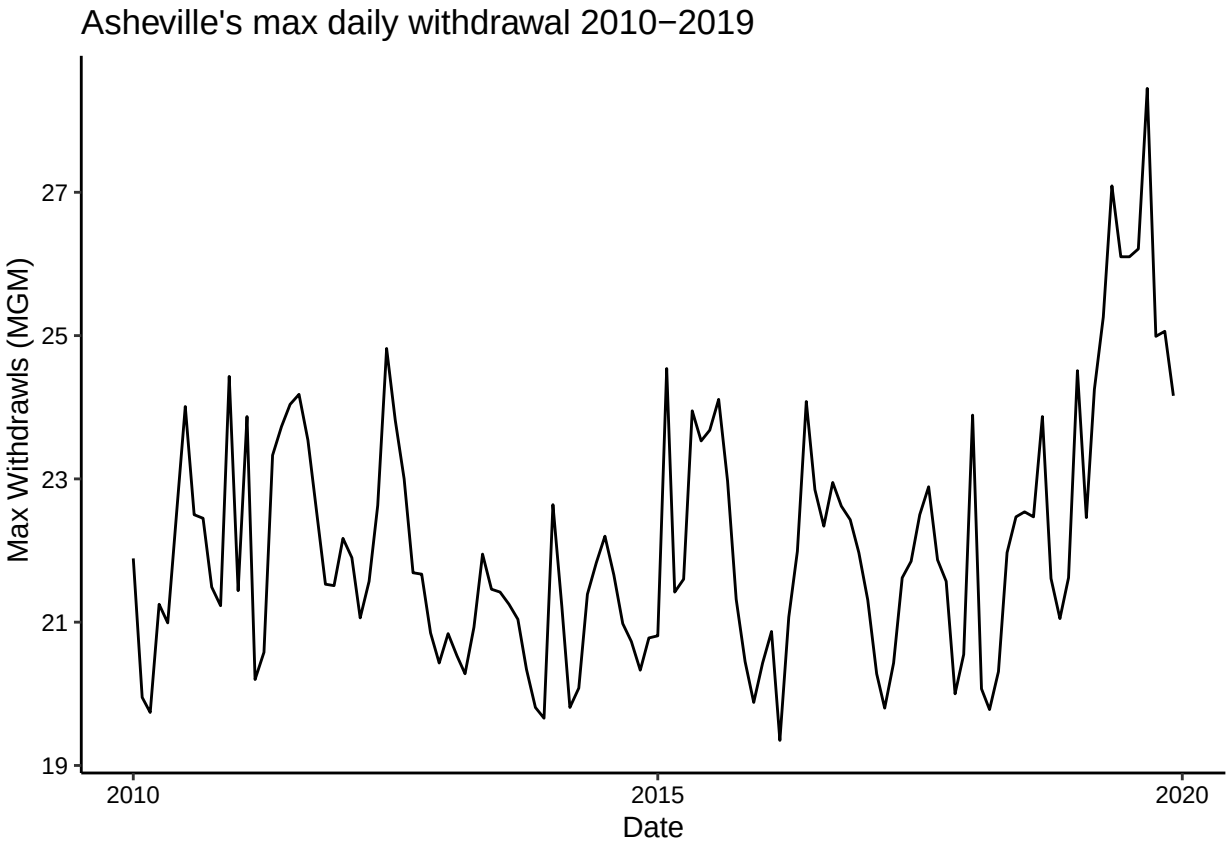
run a function over two inputs. Pipe the output of the map2() function to `bindrows()` to combine the dataframes into a single one.

```
# 9
Ash_pwsid <- "01-11-010"
Ash_year <- rep(2010:2019)
Asheville_9y_df <- map2(Ash_year, Ash_pwsid, scrape.it)
Ash_bind <- bind_rows(Asheville_9y_df)

Ash_bind$Date <- as.Date(Ash_bind$Date, format = "%Y-%M-%D")
class(Ash_bind$Date)
```

```
## [1] "Date"
```

```
Ashvile_Durham_plot_9 <- ggplot(Ash_bind, aes(x = Date,
    y = Max_Withdrawls_mgd)) + geom_line() + labs(title = paste("Asheville's max daily withdrawal 2010-
    X = "Date", y = " Max Withdrawls (MGM)")
print(Ashvile_Durham_plot_9)
```



Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time? Asheville uses more water over time. There is a peak water usage around 2017. There seems to be many seasonal and yearly fluxuuations but the max waer usage started to increase around 2017.