Assignment 10: Data Scraping

Sarah Kear

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(here)

here()
```

```
## [1] "/Users/sarah/Documents/872_EDA/EDA_Spring2024"
```

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
theURL <- 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)".

```
#3
WaterSystem <- theURL %>%
        html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
        html_text()
WaterSystem
```

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

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

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

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

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

```
MaxDayUse <- theURL %>%
        html_nodes("th~ td+ td") %>%
        html_text()
MaxDayUse
```

```
##  [1] "36.1000" "43.4200" "52.4900" "30.5000" "42.5900" "34.8800" "39.9100"
##  [8] "43.3200" "32.5300" "34.6600" "41.8000" "37.5300"
```

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.

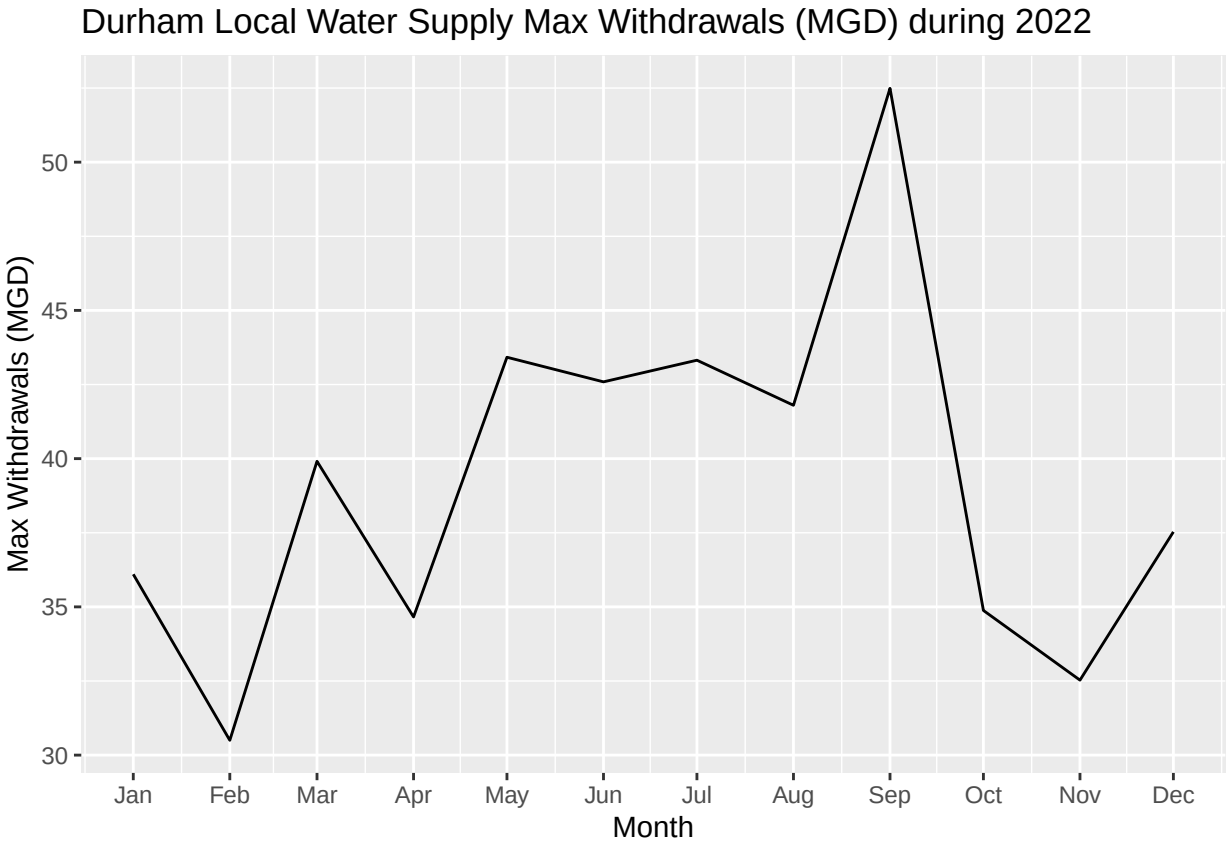
   NOTE: It's likely you won't be able to scrape the monthly widthrawal data in chronological order. You can overcome this by creating a month column manually assigning values in the order the data are scraped: "Jan", "May", "Sept", "Feb", etc... Or, you could scrape month values from the web page...

5. Create a line plot of the maximum daily withdrawals across the months for 2022

```
#4
df_withdrawals <- data.frame("WaterSystem" = rep(WaterSystem),
                             "PWSID" = rep(PWSID),
                             "Ownership" = rep(Ownership),
                             "Month" = c("Jan", "May", "Sept",
                             "Feb","Jun", "Oct",
                             "Mar", "Jul", "Nov",
                             "Apr", "Aug", "Dec"),
                             "Year" = rep(2022,12),
                             "Max_Withdrawals_mgd" = as.numeric(MaxDayUse)) %>%
        mutate(Date = my(paste(Month,"-",Year)))
df_withdrawals
```

```
##    WaterSystem     PWSID    Ownership Month Year Max_Withdrawals_mgd       Date
## 1       Durham 03-32-010 Municipality   Jan 2022               36.10 2022-01-01
## 2       Durham 03-32-010 Municipality   May 2022               43.42 2022-05-01
## 3       Durham 03-32-010 Municipality  Sept 2022               52.49 2022-09-01
## 4       Durham 03-32-010 Municipality   Feb 2022               30.50 2022-02-01
## 5       Durham 03-32-010 Municipality   Jun 2022               42.59 2022-06-01
## 6       Durham 03-32-010 Municipality   Oct 2022               34.88 2022-10-01
## 7       Durham 03-32-010 Municipality   Mar 2022               39.91 2022-03-01
## 8       Durham 03-32-010 Municipality   Jul 2022               43.32 2022-07-01
## 9       Durham 03-32-010 Municipality   Nov 2022               32.53 2022-11-01
## 10      Durham 03-32-010 Municipality   Apr 2022               34.66 2022-04-01
## 11      Durham 03-32-010 Municipality   Aug 2022               41.80 2022-08-01
## 12      Durham 03-32-010 Municipality   Dec 2022               37.53 2022-12-01
```

```
#5
plot1 <- df_withdrawals %>%
        ggplot(aes(x=Date,
                   y=Max_Withdrawals_mgd)) +
        geom_line() +
         scale_x_date(date_labels = "%b", date_breaks = "1 month") +
        ylab("Max Withdrawals (MGD)") +
        xlab("Month") +
        ggtitle(paste(WaterSystem,"Local Water Supply Max Withdrawals (MGD) during 2022"))
plot1
```



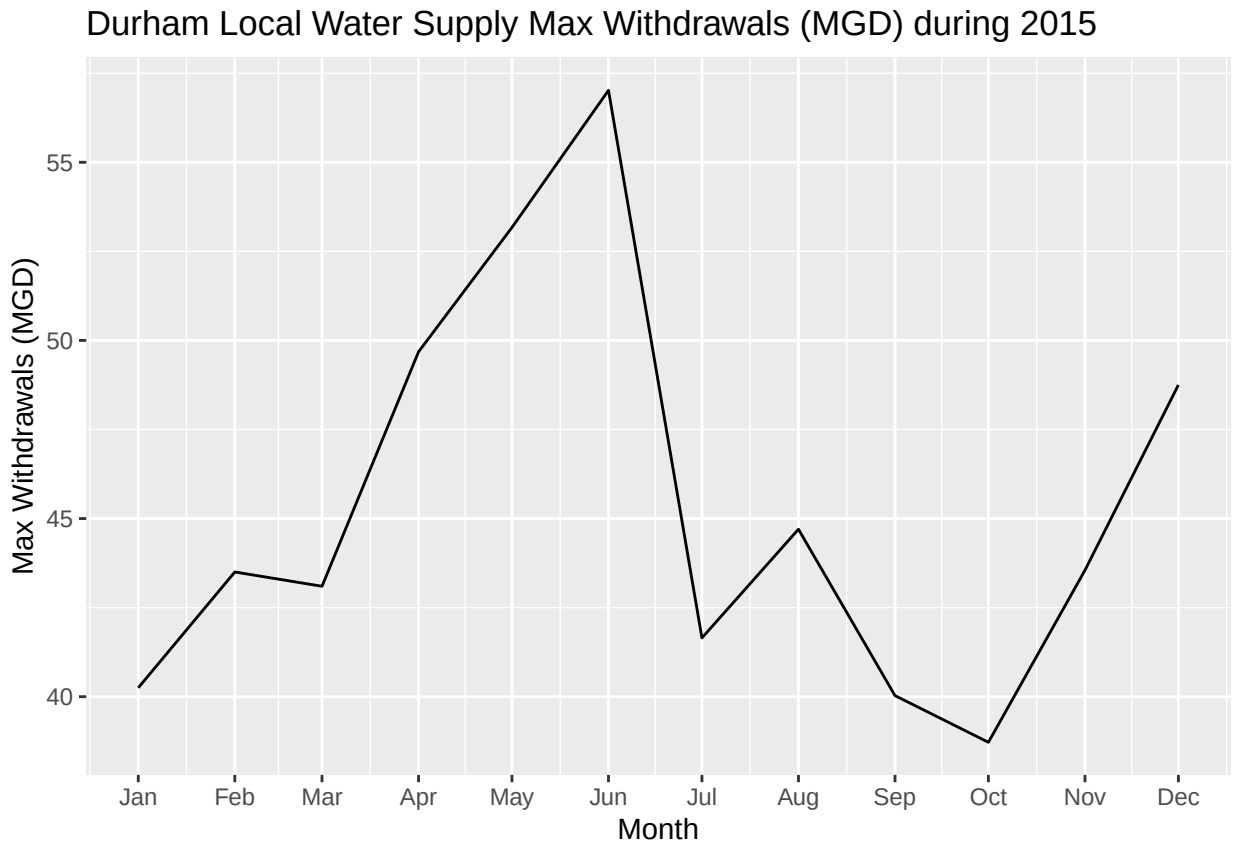
6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

```
#6.
water.data.scrape <- function(the_PWSID, the_year){

        #Retrieving web contents
        the_website <- read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?',
                                        'pwsid=', the_PWSID, '&year=', the_year))

        #Scraping the elements from webpage
        WaterSystem <- the_website %>%
                html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
                html_text()
        PWSID <- the_website %>%
                html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
                html_text()
        Ownership <- the_website %>%
                html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%
                html_text()
        MaxDayUse <- the_website %>%
                html_nodes("th~ td+ td") %>%
                html_text()
```

```
        #Convert to a dataframe
        df_withdrawals <- data.frame("WaterSystem" = rep(WaterSystem),
                                     "PWSID" = rep(PWSID),
                                     "Ownership" = rep(Ownership),
                                     "Month" = c("Jan", "May", "Sept",
                                                 "Feb","Jun", "Oct",
                                                 "Mar", "Jul", "Nov",
                                                 "Apr", "Aug", "Dec"),
                                     "Year" = rep(the_year,12),
                                     "Max_Withdrawals_mgd" = as.numeric(MaxDayUse)) %>%
                mutate(Date = my(paste(Month,"-",Year)))
        df_withdrawals

        Sys.sleep(5) #uncomment this if you are doing bulk scraping!

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

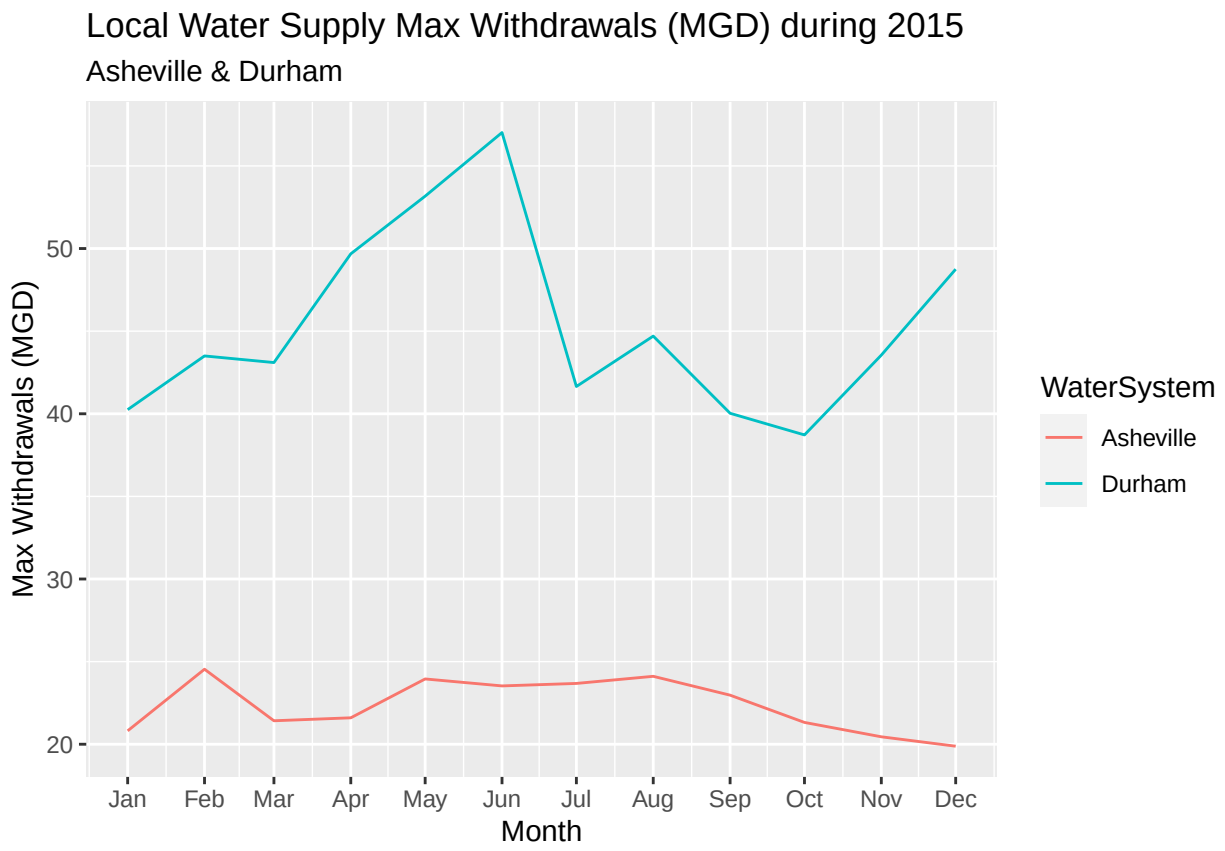
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID=‘03-32-010’) for each month in 2015

```
#7
q7_df <- water.data.scrape('03-32-010',2015)
view(q7_df)

plot2 <- q7_df %>%
        ggplot(aes(x=Date,
                   y=Max_Withdrawals_mgd)) +
        geom_line() +
         scale_x_date(date_labels = "%b", date_breaks = "1 month") +
        ylab("Max Withdrawals (MGD)") +
        xlab("Month") +
        ggtitle(paste(WaterSystem,"Local Water Supply Max Withdrawals (MGD) during 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 Asheville's to Durham's water withdrawals.

```
#8
q8_df <- water.data.scrape('01-11-010',2015)
view(q8_df)

Ashville.Durham.2015_df <- bind_rows(q7_df, q8_df)

plot3 <- Ashville.Durham.2015_df %>%
        ggplot(aes(x=Date,
                   y=Max_Withdrawals_mgd,
                   color = WaterSystem)) +
        geom_line() +
         scale_x_date(date_labels = "%b", date_breaks = "1 month") +
        ylab("Max Withdrawals (MGD)") +
        xlab("Month") +
        ggtitle("Local Water Supply Max Withdrawals (MGD) during 2015") +
        labs(subtitle = "Asheville & Durham")
plot3
```



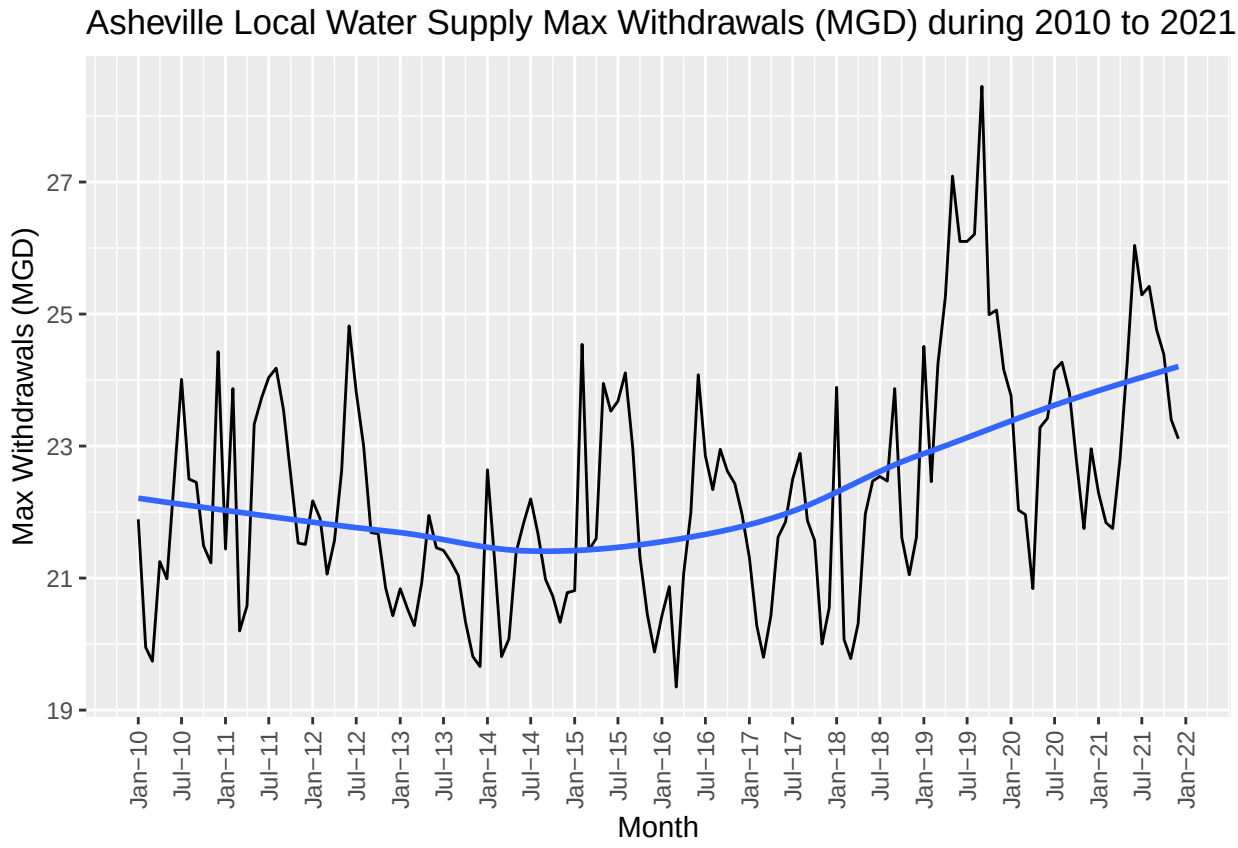
9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2021.Add a smoothed line to the plot (method = 'loess').

   TIP: See Section 3.2 in the "10_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
the_years = rep(2010:2021)
my_PWSID = '01-11-010'

#Use lapply to apply the scrape function
the_dfs <- lapply(X = the_years,
                  FUN = water.data.scrape,
                  the_PWSID=my_PWSID)

Asheville_df <- bind_rows(the_dfs)

plot4 <- Asheville_df %>%
        ggplot(aes(x=Date,
                   y=Max_Withdrawals_mgd)) +
        geom_line() +
        geom_smooth(method="loess",se=FALSE) +
        scale_x_date(breaks = seq(as.Date("2010-01-01"),
                                  as.Date("2022-01-01"),
```

```
                                  by="6 months"),
                     date_labels = "%b-%y") +
        ylab("Max Withdrawals (MGD)") +
        xlab("Month") +
        ggtitle("Asheville Local Water Supply Max Withdrawals (MGD) during 2010 to 2021") +
        theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

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: Between 2010 and 2016, there seems to be a slight decrease in water usage. But then between 2016 and 2021, there is a staeady increase in water usage. >