Assignment 10: Data Scraping

Student Name

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)
library(here)
here()
```

```
## [1] "C:/Users/purec/Documents/Duke/Fall_2023/EDA/EDE_Fall2023"
```

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.)

```r
#2
website <- 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)".

```r
#3
system_name <- website %>%
        html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>%
        html_text()

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

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

max_mgd <- website %>%
        html_nodes('th~ td+ td') %>%
        html_text()

mgd_months <- website %>%
        html_nodes('.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...

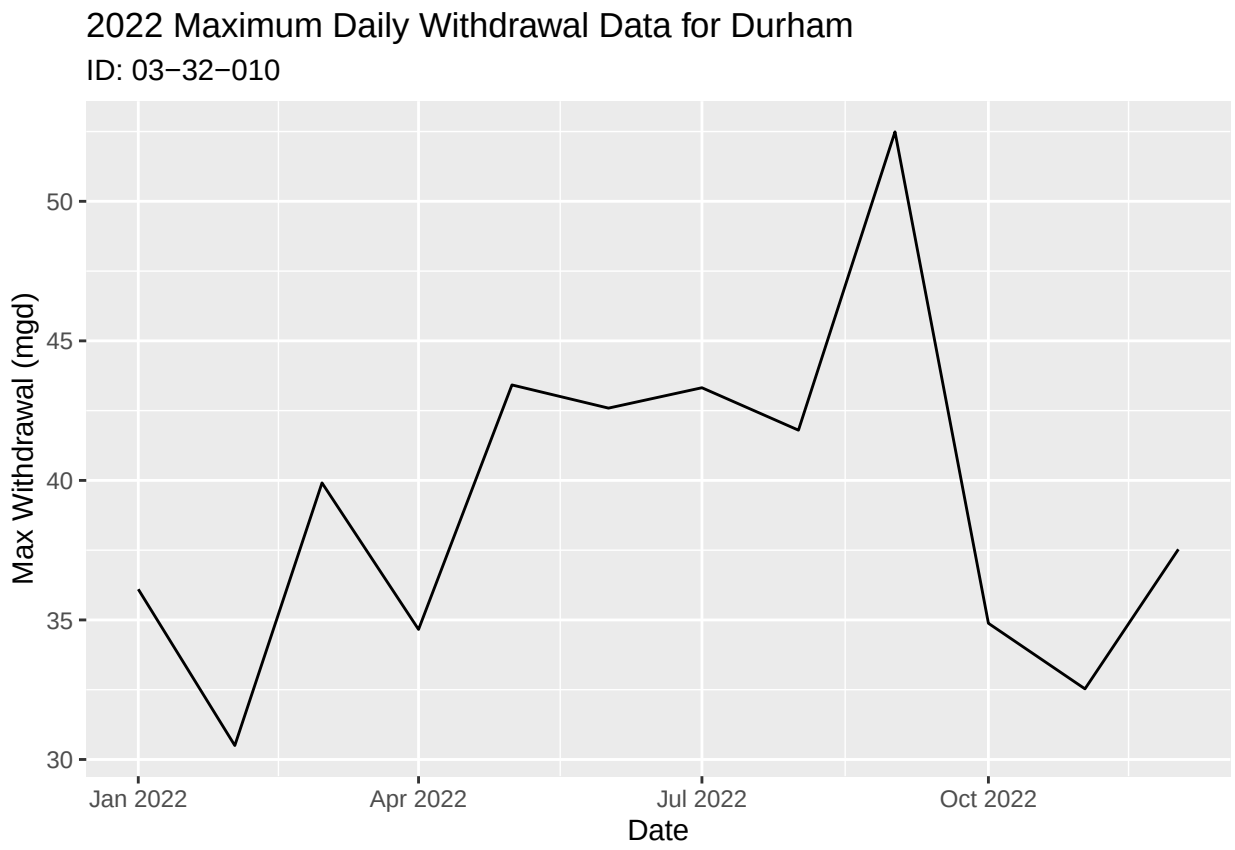
5. Create a line plot of the maximum daily withdrawals across the months for 2022

```
#4
max_mgd_df <- data.frame("Ownership" = rep(ownership, 12),
                         "County" = rep(system_name, 12),
                         "PWSID" = rep(pwsid, 12),
                         "Month" = mgd_months,
                         "Year" = rep(2022, 12),
                         "Max_Withdrawals_mgd" = as.numeric(max_mgd))

max_mgd_df <- max_mgd_df %>%
        mutate("MonthYear" = my(paste(Month,"-",Year))) %>%
        arrange(MonthYear)

#5
max_mgd_lineplot <- ggplot(max_mgd_df,aes(x=MonthYear,y=Max_Withdrawals_mgd)) +
  geom_line()+
  labs(title = paste("2022 Maximum Daily Withdrawal Data for",system_name),
       subtitle = paste('ID:',pwsid),
       y="Max Withdrawal (mgd)",
       x="Date")

max_mgd_lineplot
```



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_w_pwsid <- function(the_year, the_pwsid){

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


  system_name_tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'
  ownership_tag <- 'div+ table tr:nth-child(2) td:nth-child(4)'
  max_mgd_tag <- 'th~ td+ td'

  #Scrape the data items
  systemname <- the_website %>% html_nodes(system_name_tag) %>% html_text()
  owner <- the_website %>%html_nodes(ownership_tag) %>% html_text()
  maxmgd <- the_website %>% html_nodes(max_mgd_tag) %>% html_text()

  #Convert to a dataframe
  maxwithdrawals <- data.frame("Ownership" = rep(owner, 12),
                           "County" = rep(systemname, 12),
                           "PWSID" = rep(the_pwsid, 12),
                           "Month" = mgd_months,
                           "Year" = rep(the_year, 12),
                           "Max_Withdrawals_mgd" = as.numeric(maxmgd)) %>%
          mutate("MonthYear" = my(paste(Month,"-",Year))) %>%
          arrange(MonthYear)

  Sys.sleep(1)

  return(maxwithdrawals)
}
```

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

```
#7
durham15_maxwithdrawals <- scrape_w_pwsid('2015', '03-32-010')
durham15_maxwithdrawals
```

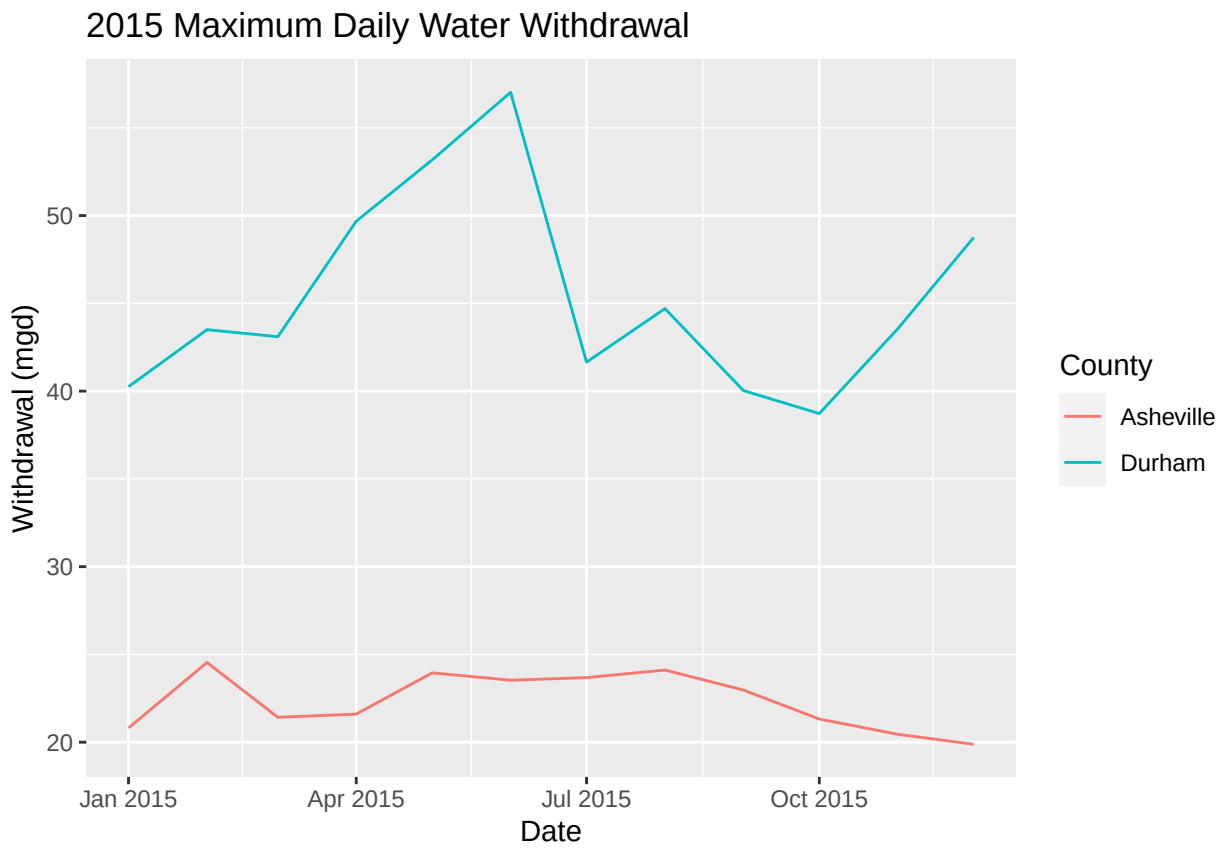
```
##       Ownership County     PWSID Month Year Max_Withdrawals_mgd  MonthYear
## 1  Municipality Durham 03-32-010   Jan 2015               40.25 2015-01-01
## 2  Municipality Durham 03-32-010   Feb 2015               43.50 2015-02-01
## 3  Municipality Durham 03-32-010   Mar 2015               43.10 2015-03-01
## 4  Municipality Durham 03-32-010   Apr 2015               49.68 2015-04-01
## 5  Municipality Durham 03-32-010   May 2015               53.17 2015-05-01
## 6  Municipality Durham 03-32-010   Jun 2015               57.02 2015-06-01
## 7  Municipality Durham 03-32-010   Jul 2015               41.65 2015-07-01
## 8  Municipality Durham 03-32-010   Aug 2015               44.70 2015-08-01
## 9  Municipality Durham 03-32-010   Sep 2015               40.03 2015-09-01
## 10 Municipality Durham 03-32-010   Oct 2015               38.72 2015-10-01
## 11 Municipality Durham 03-32-010   Nov 2015               43.55 2015-11-01
## 12 Municipality Durham 03-32-010   Dec 2015               48.75 2015-12-01
```

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
asheville15_maxwithdrawals <- scrape_w_pwsid('2015', '01-11-010')

ash_drm <- rbind(asheville15_maxwithdrawals, durham15_maxwithdrawals)

ash_drm_15_plot <- ggplot(ash_drm, aes(x=MonthYear, y=Max_Withdrawals_mgd, color=County))+
        geom_line()+
        labs(title = "2015 Maximum Daily Water Withdrawal",
             x = "Date",
             y = "Withdrawal (mgd)")
ash_drm_15_plot
```



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.

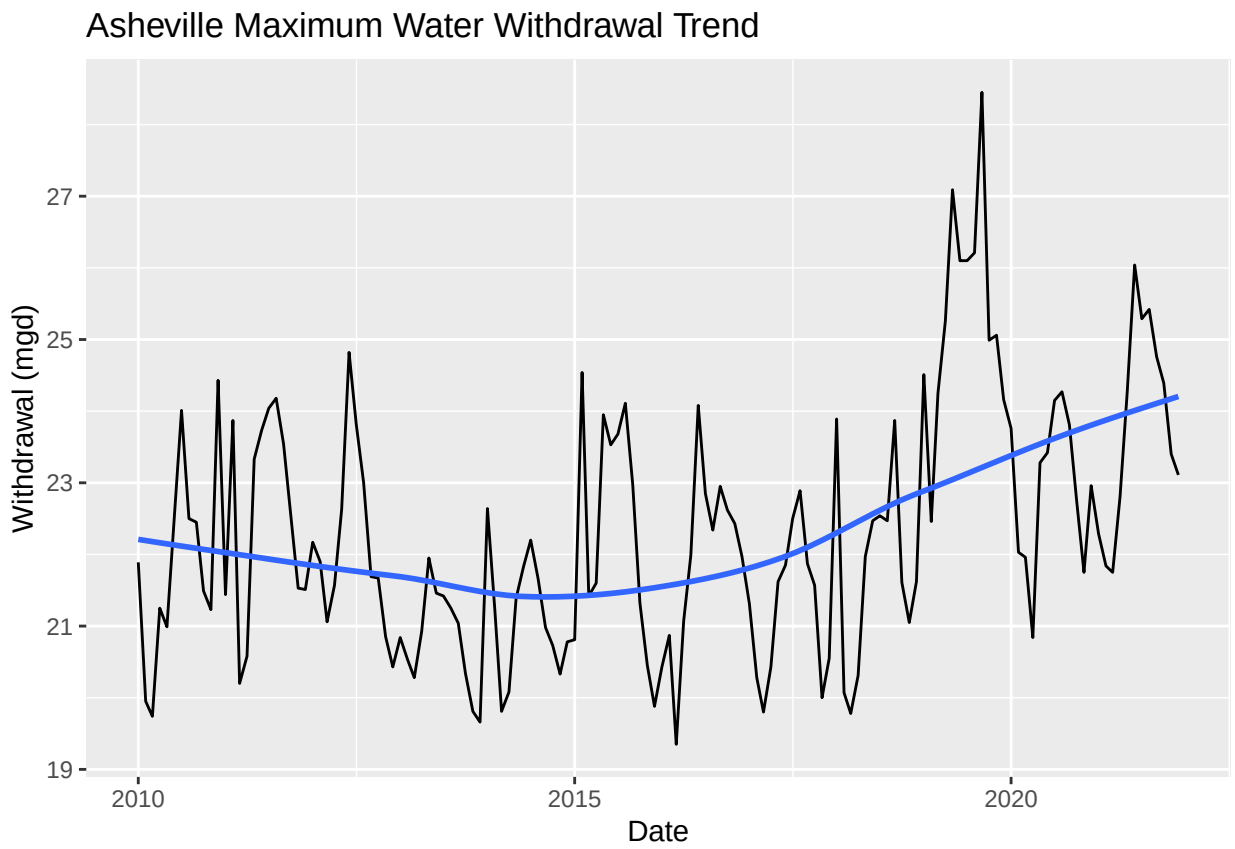
```
#9
years <- c(2010:2021)
pwsids_ash <- rep("01-11-010", length(years))

asheville_maxmgd <- map2(years, pwsids_ash, scrape_w_pwsid) %>%
        bind_rows()

ash_1021_plot <- ggplot(asheville_maxmgd, aes(x=MonthYear, y=Max_Withdrawals_mgd))+
        geom_line()+
        geom_smooth(method='loess', se=FALSE)+
        labs(title = "Asheville Maximum Water Withdrawal Trend",
             x = "Date",
             y = "Withdrawal (mgd)")

ash_1021_plot
```

```
## `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: Not definitively– there seems to be a slight decrease until 2015, and then a stronger increase in usage from then on. This probably relates to Asheville's increasing popularity as a desired city to move to.