Assignment 10: Data Scraping

Shiqi Zheng

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/shiqizheng/Desktop/ENV872/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.)

```
#2
LWSP <- read_html(
  'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2022')
LWSP
```

```
## {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
- 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
name <- LWSP %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()
name
```

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

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

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

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

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

```
Maximum.Day.Use <- LWSP %>%
  html_nodes("th~ td+ td") %>%
  html_text()
Maximum.Day.Use
```

```
##  [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.

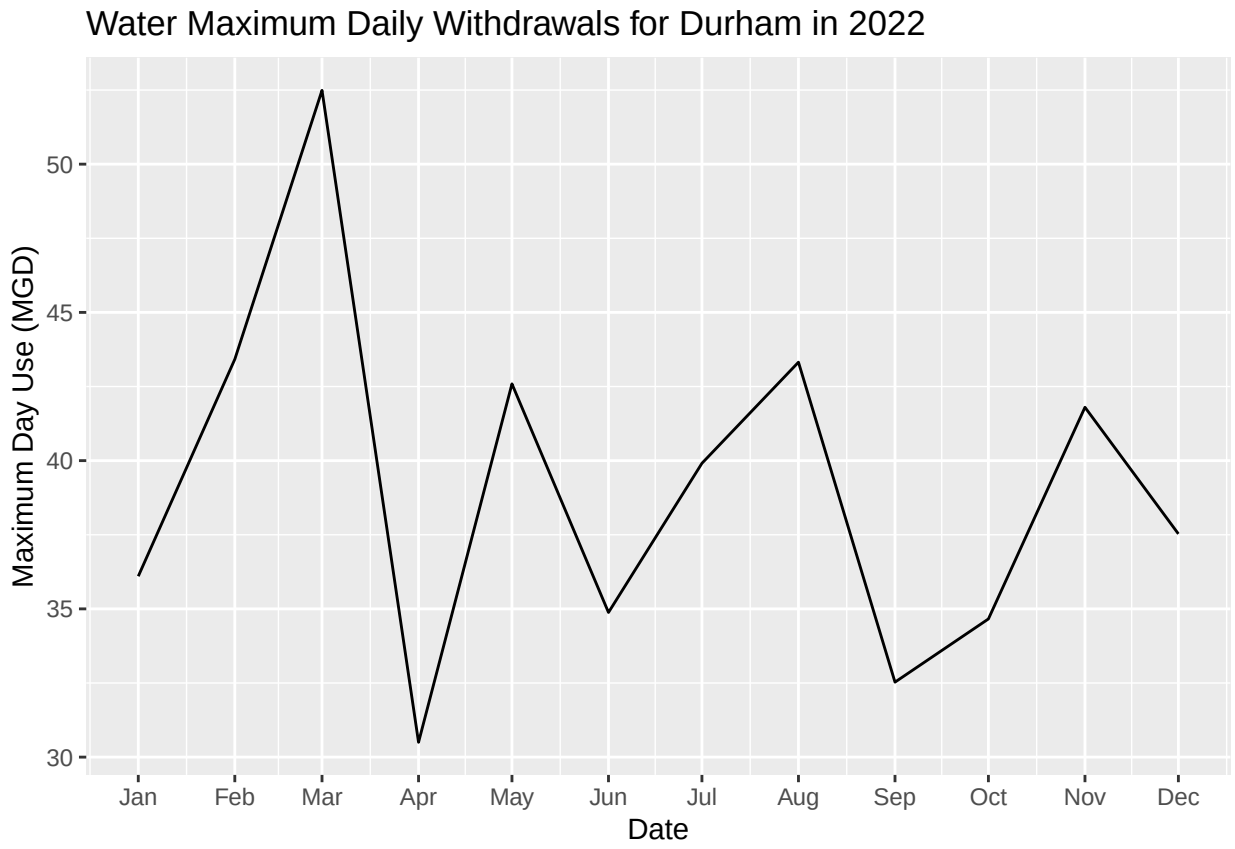
   NOTE: It's likely you won't be able to scrape the monthly widthrawal data in chronological order. You can overcome this by creating a month column manually assigning values in the order the data are scraped: "Jan", "May", "Sept", "Feb", etc... Or, you could scrape month values from the web page...

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

```
#4
df <- data.frame(
  "WaterSystemName" = rep(name,12),
  "PWSID" = rep(PWSID, 12),
  "Ownership" = rep(ownership,12),
  "MaximumDayUse" = as.numeric(Maximum.Day.Use),
  "Month" = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul",
              "Aug", "Sep", "Oct", "Nov", "Dec"),
  "Year" = rep(2022,12)
) %>%
  mutate(Date = my(paste(Month,"-",Year)))
head(df)
```

```
##   WaterSystemName     PWSID    Ownership MaximumDayUse Month Year       Date
## 1          Durham 03-32-010 Municipality         36.10   Jan 2022 2022-01-01
## 2          Durham 03-32-010 Municipality         43.42   Feb 2022 2022-02-01
## 3          Durham 03-32-010 Municipality         52.49   Mar 2022 2022-03-01
## 4          Durham 03-32-010 Municipality         30.50   Apr 2022 2022-04-01
## 5          Durham 03-32-010 Municipality         42.59   May 2022 2022-05-01
## 6          Durham 03-32-010 Municipality         34.88   Jun 2022 2022-06-01
```

```
# ?? a row; modify reflect year

#5
ggplot(df,aes(x=Date,y=MaximumDayUse)) +
  geom_line() +
   scale_x_date(date_labels = "%b", date_breaks = "1 month") +
  labs(title = "Water Maximum Daily Withdrawals for Durham in 2022",
       y="Maximum Day Use (MGD)",
       x="Date")
```



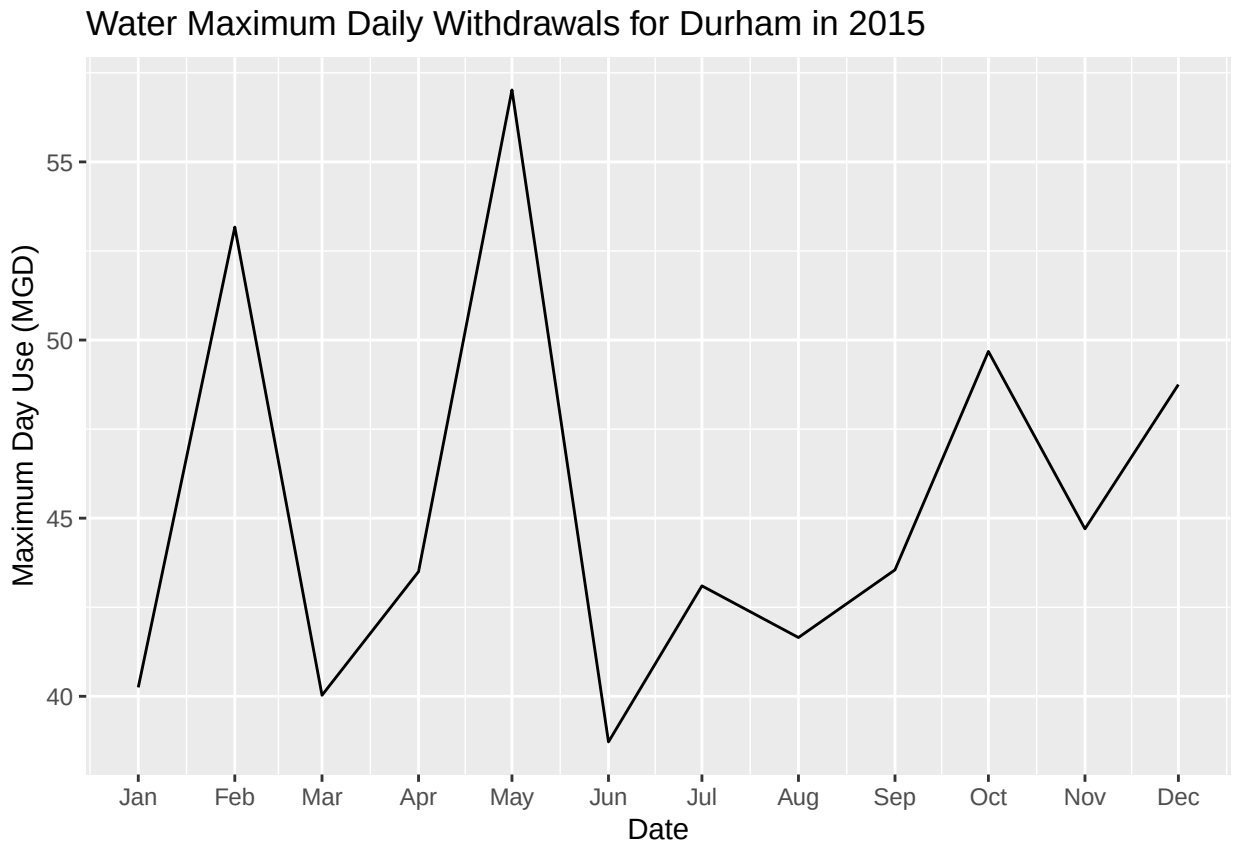
6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

```
#6.
scrape.it <- function(year, PWSID){
   LWSP <- read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',
                                PWSID, '&year=',year))
#scape data
name <- LWSP %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>% html_text()
PWSID <- LWSP %>% html_nodes("td tr:nth-child(1) td:nth-child(5)") %>% html_text()
ownership <- LWSP %>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>% html_text()
Maximum.Day.Use <- LWSP %>% html_nodes("th~ td+ td") %>% html_text()

# create dataframe
df <- data.frame(
  "WaterSystemName" = rep(name,12),
  "PWSID" = rep(PWSID, 12),
  "Ownership" = rep(ownership,12),
  "MaximumDayUse" = as.numeric(Maximum.Day.Use),
  "Month" = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul",
              "Aug", "Sep", "Oct", "Nov", "Dec"),
  "Year" = rep(year,12)
```

```
) %>%
  mutate(Date = my(paste(Month,"-",Year)))

return(df)
}
```

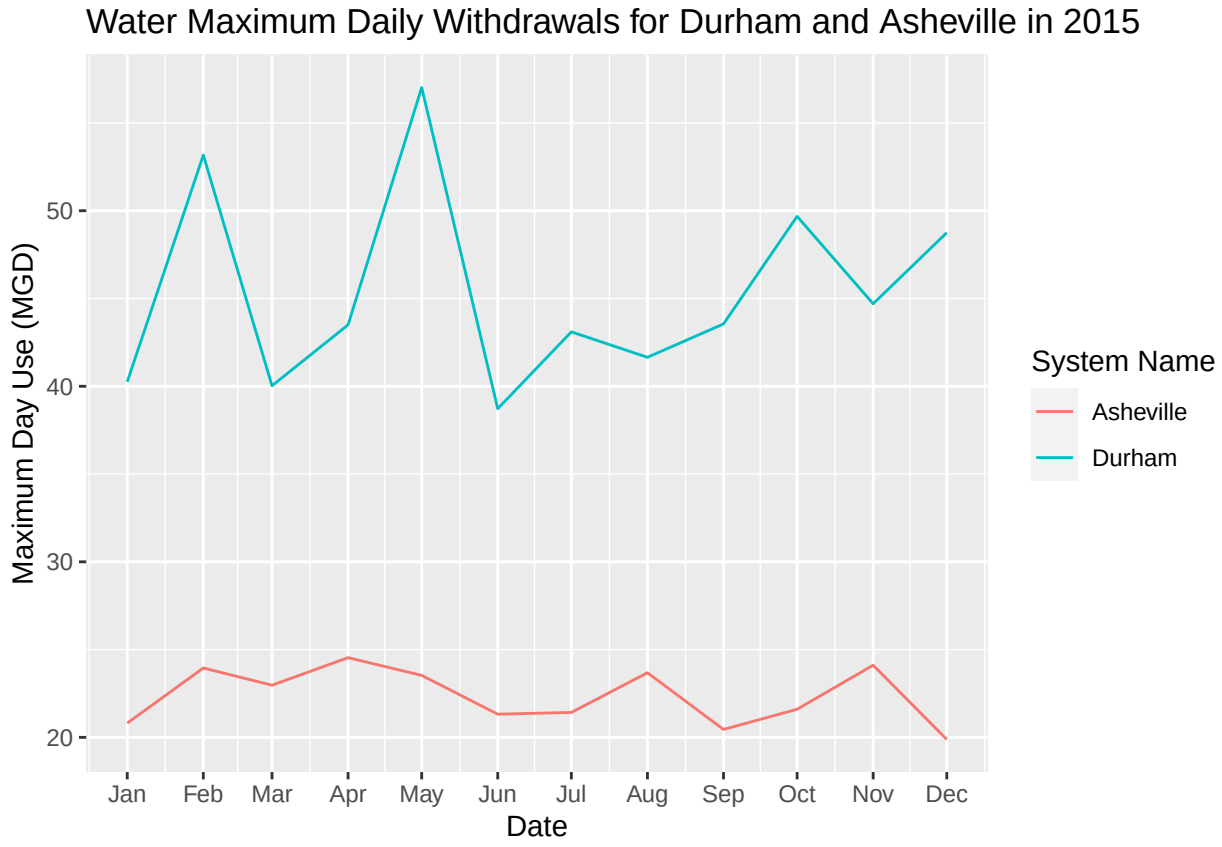
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID='03-32-010') for each month in 2015

```
#7
the_df <- scrape.it(2015,'03-32-010')
ggplot(the_df,aes(x=Date,y=MaximumDayUse)) +
  geom_line() +
  scale_x_date(date_labels = "%b", date_breaks = "1 month") +
  labs(title = "Water Maximum Daily Withdrawals for Durham in 2015",
       y="Maximum Day Use (MGD)",
       x="Date")
```



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.

```r
#8
Asheville_df <- scrape.it(2015,'01-11-010')
combine_df <- rbind(Asheville_df,the_df)
ggplot(combine_df,aes(x=Date,y=MaximumDayUse,color = WaterSystemName)) +
  geom_line() +
  scale_x_date(date_labels = "%b", date_breaks = "1 month") +
  labs(title = "Water Maximum Daily Withdrawals for Durham and Asheville in 2015",
       y="Maximum Day Use (MGD)",
       x="Date",
       color ="System Name")
```



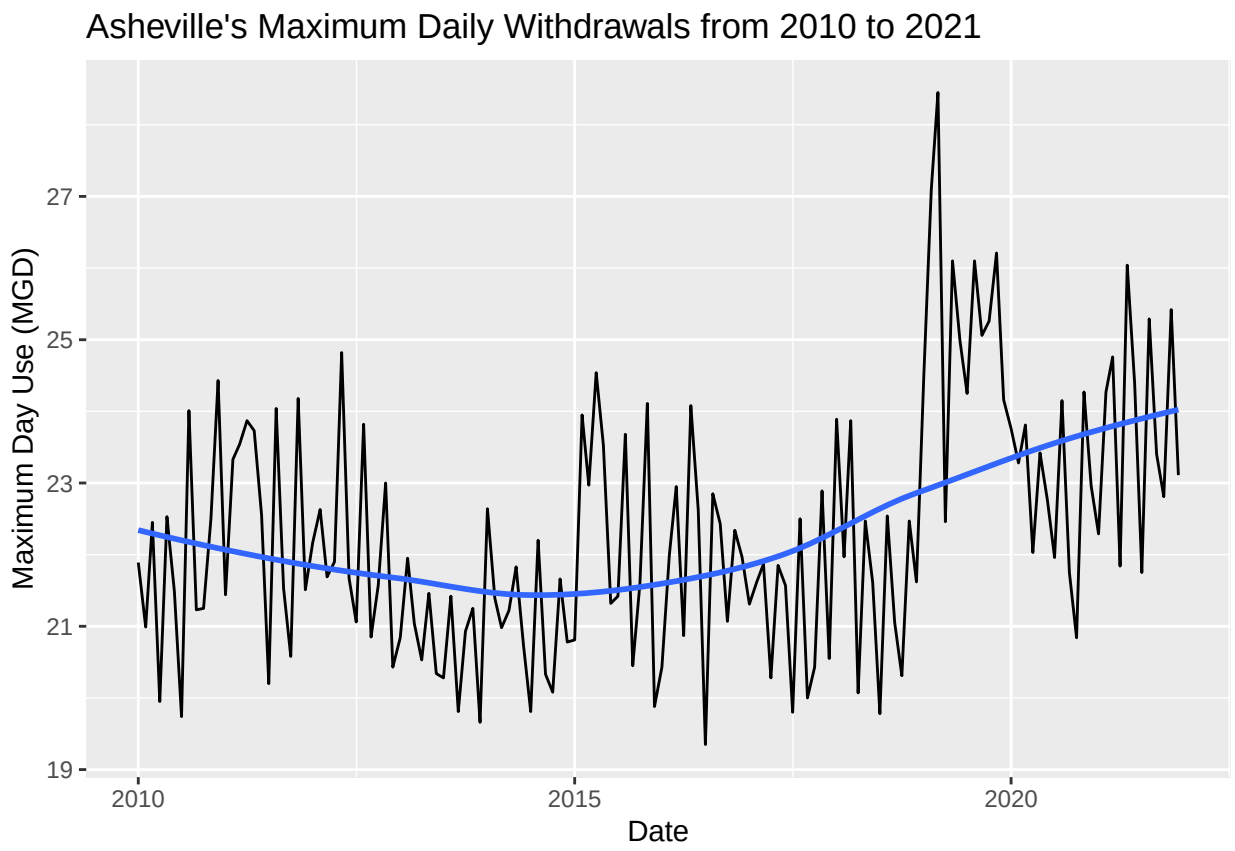
9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2021.Add a smoothed line to the plot (method = 'loess').

   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.

```r
#9
the_years <- seq(2010,2021)
the_PWSID <-  rep('01-11-010',length(the_years))

dfs_Asheville <- map2(the_years, the_PWSID , scrape.it) %>% bind_rows()
ggplot(dfs_Asheville, aes(x = Date, y = MaximumDayUse)) +
```

```
  geom_line() +
  geom_smooth(method = 'loess', se = FALSE) +
  labs(title = "Asheville's Maximum Daily Withdrawals from 2010 to 2021",
       y = "Maximum Day Use (MGD)",
       x = "Date")
```

```
## `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: From 2010 to 2015, Asheville has a decreasing trend in water usage. From 2015 to 2020, Asheville has a increasing trend in water usage through time.