Assignment 10: Data Scraping

Yin-Chia Yang

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)
library(ggplot2)
library(dplyr)

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

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
the_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)".

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

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

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

the_MGD <- the_website %>% html_nodes("th~ td+ td") %>% html_text()

the_month <- the_website %>% html_nodes(".fancy-table:nth-child(31) tr+ tr th") %>% html_text()
```

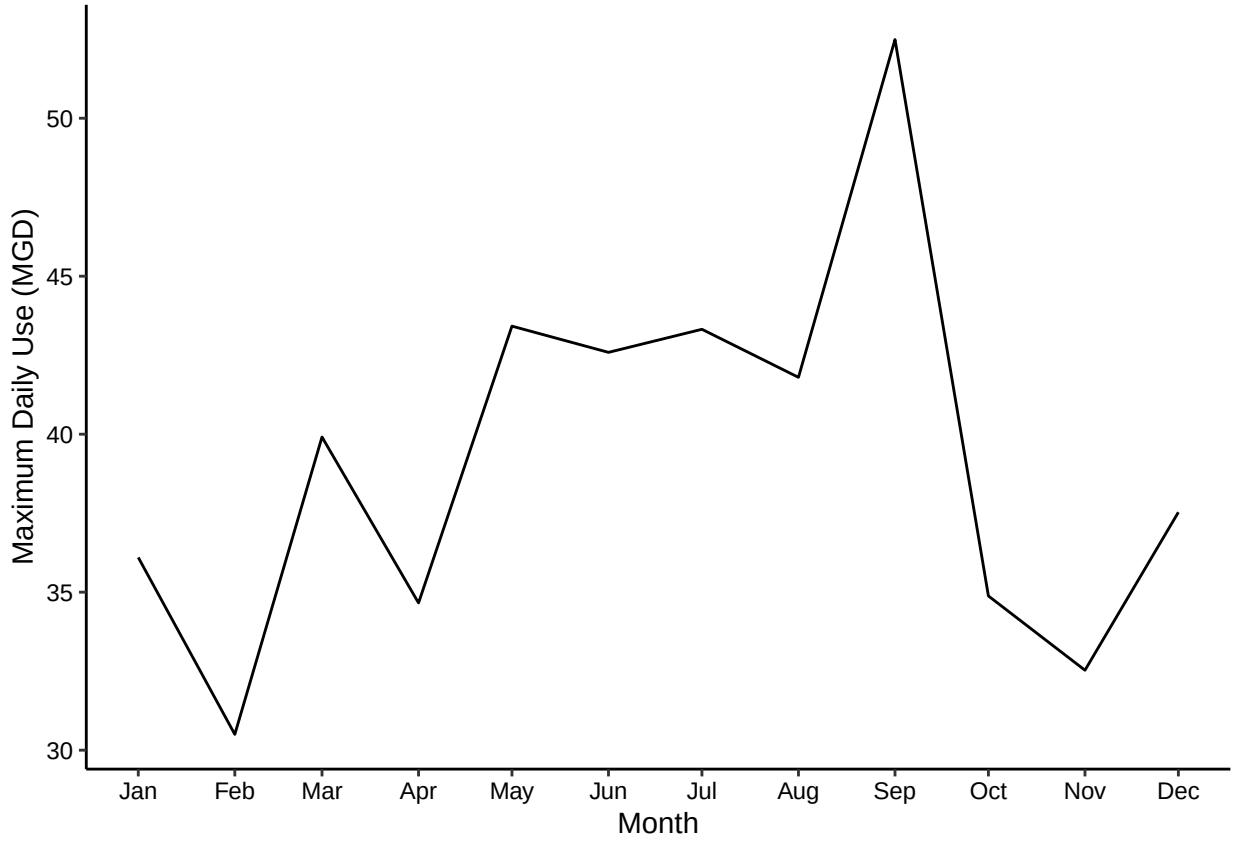
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   > 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 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

```r
#4
the_df <- data.frame(
  Year = '2022',
  Month = the_month,
  water_system_name = the_water_sys_nm,
  PWSID = the_PWSID,
  Ownership = the_Ownership,
  max_daily_use = as.numeric(the_MGD)) %>%
  mutate(
    Date = my(paste(Month, "-", Year))
  ) %>%
  arrange(Date)

#5
ggplot(the_df, aes(x= Date, y= max_daily_use)) +
  geom_line(aes()) +
  scale_x_date(date_breaks = "1 month", date_labels = '%b') +
  labs(x='Month', y='Maximum Daily Use (MGD)')
```

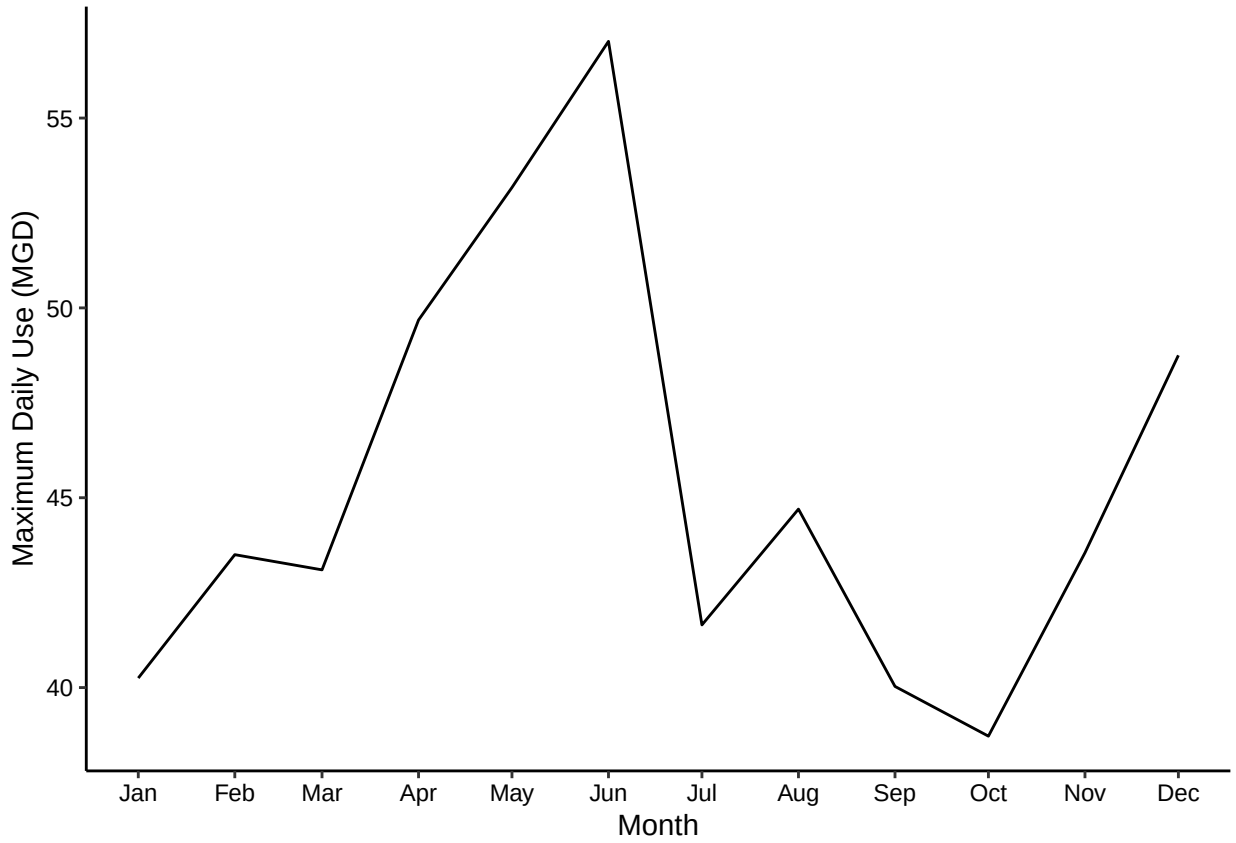


6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

```r
#6.
scrape.it <- function(the_year, the_PWSID){

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

  the_water_sys_nm <- the_website %>%
    html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>% html_text()
  the_PWSID <- the_website %>%
    html_nodes("td tr:nth-child(1) td:nth-child(5)") %>% html_text()
  the_Ownership <- the_website %>%
    html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%  html_text()
  the_MGD <- the_website %>% html_nodes("th~ td+ td") %>% html_text()

  the_df <- data.frame(
  Year = the_year,
  Month = the_month,
  water_system_name = the_water_sys_nm,
  PWSID = the_PWSID,
  Ownership = the_Ownership,
  max_daily_use = as.numeric(the_MGD)) %>%
   mutate(
    Date = my(paste(Month, "-", Year))
  ) %>%
  arrange(Date)
  Sys.sleep(1)
  return(the_df)
}
```

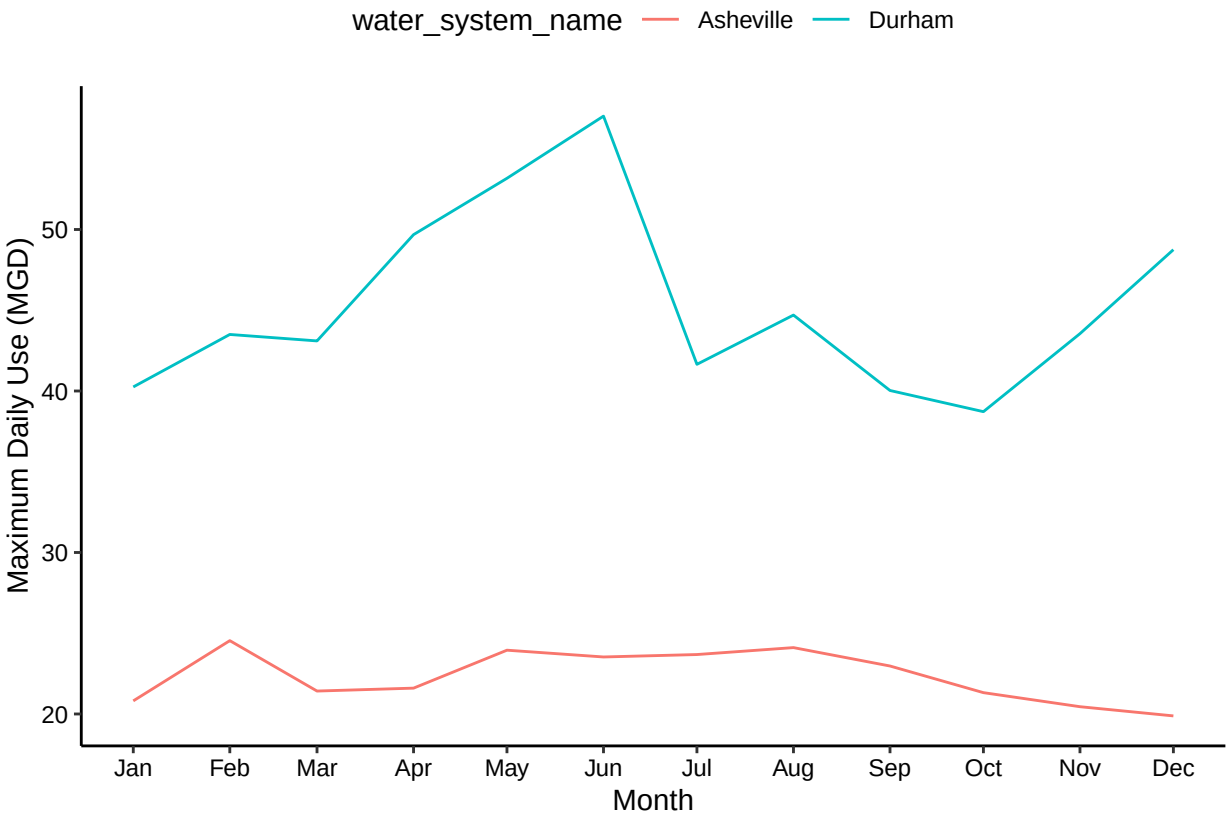
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID=‘03-32-010’) for each month in 2015

```r
#7
the_df_durham <- scrape.it(2015, '03-32-010')

ggplot(the_df_durham, aes(x= Date, y= max_daily_use)) +
  geom_line(aes()) +
  scale_x_date(date_breaks = "1 month", date_labels = '%b') +
  labs(x='Month', y='Maximum Daily Use (MGD)')
```



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
the_df_asheville <- scrape.it(2015, '01-11-010')

the_df_durham_aseville <- rbind.data.frame(the_df_durham, the_df_asheville)

ggplot(the_df_durham_aseville, aes(x= Date, y= max_daily_use)) +
  geom_line(aes(color= water_system_name)) +
  scale_x_date(date_breaks = "1 month", date_labels = '%b') +
  labs(x='Month', y='Maximum Daily Use (MGD)')
```

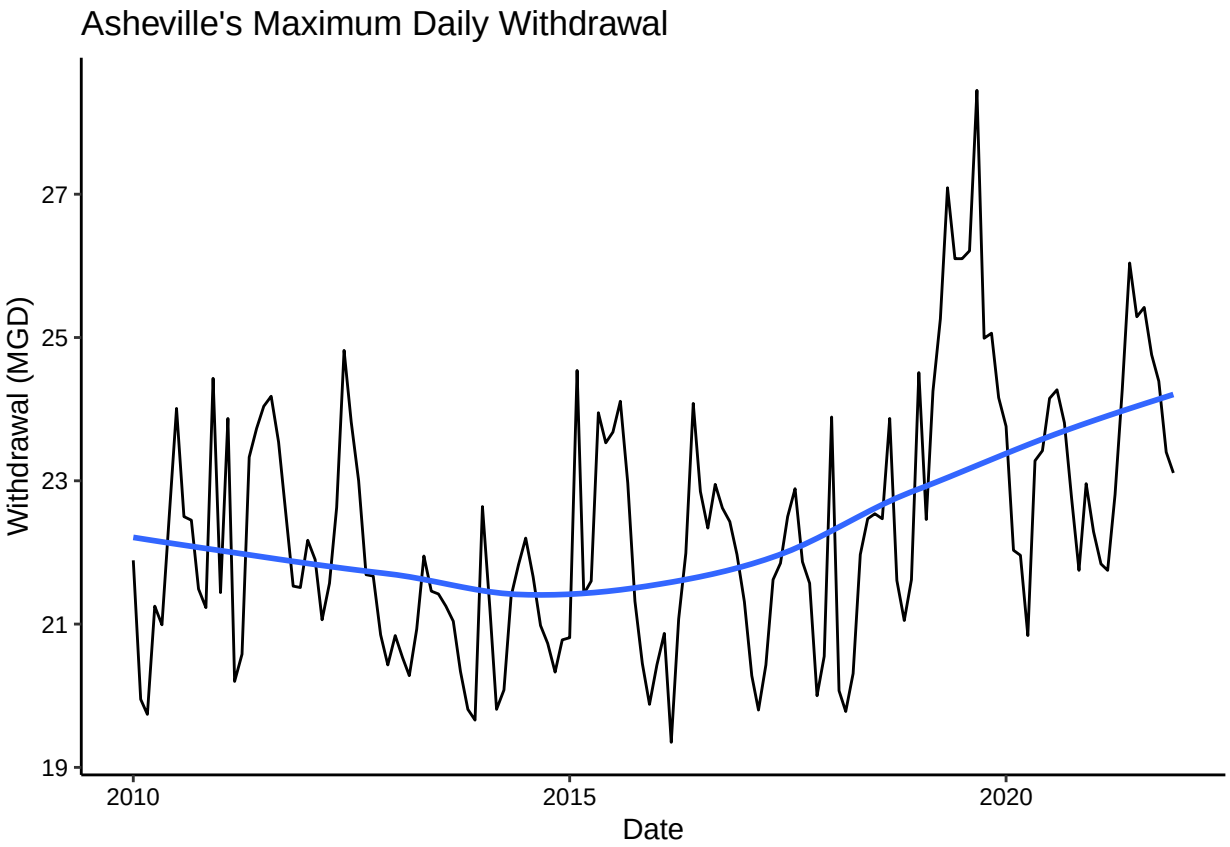


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2021.Add a smoothed line to the plot (method = 'loess').

   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)
the_asheville_trend_df <- lapply(X = the_years,
                           FUN = scrape.it,
                           the_PWSID ='01-11-010')
the_asheville_trend_df <- bind_rows(the_asheville_trend_df)

ggplot(the_asheville_trend_df, aes(x= Date, y= max_daily_use)) +
  geom_line(aes()) +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = "Asheville's Maximum Daily Withdrawal",
       x="Date",
       y="Withdrawal (MGD)"
       )
```

```
## `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: The water usage in Asheville decreased from 2011 to 2015, and increased from 2015 to 2021. >