Assignment 09: Data Scraping

Cal Oakley

Total points:

OVERVIEW

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

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Fay_09_Data_Scraping.Rmd") prior to submission.

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] "K:/GradSchool/Spring2022/EnvironmentalDataAnalytics/Environmental_Data_Analytics_2022/Assignmen
```

```
library(tidyverse)
library(rvest)
library(lubridate)

CalsTheme <- theme_classic(base_size = 16) +
  theme(axis.text = element_text(color = "gray"), legend.position = "left",
        legend.justification = 2)
theme_set(CalsTheme)
```

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham's 2019 Municipal Local Water Supply Plan (LWSP):

- Navigate to https://www.ncwater.org/WUDC/app/LWSP/search.php
- Change the date from 2020 to 2019 in the upper right corner.
- 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=2020

Indicate this website as the as the URL to be scraped. (In other words, read the contents into an `rvest` webpage object.)

```
#2
theWebsite <- read_html(
  'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2020')
```

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:
- Average Daily Use (MGD) - for each month

In the code chunk below scrape these values, assigning them to three 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, with the first value being 36.0100.

```
#3
water.system.name <- theWebsite %>%
  html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>% html_text()
pwsid <- theWebsite %>%
  html_nodes('td tr:nth-child(1) td:nth-child(5)') %>% html_text()
ownership <- theWebsite %>%
  html_nodes('div+ table tr:nth-child(2) td:nth-child(4)') %>% html_text()
max.withdrawals.mgd <- theWebsite %>%
  html_nodes('th~ td+ td') %>% 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 order. You can overcome this by creating a month column in the same order the data are scraped: Jan, May, Sept, Feb, etc. . .

5. Plot the max daily withdrawals across the months for 2020

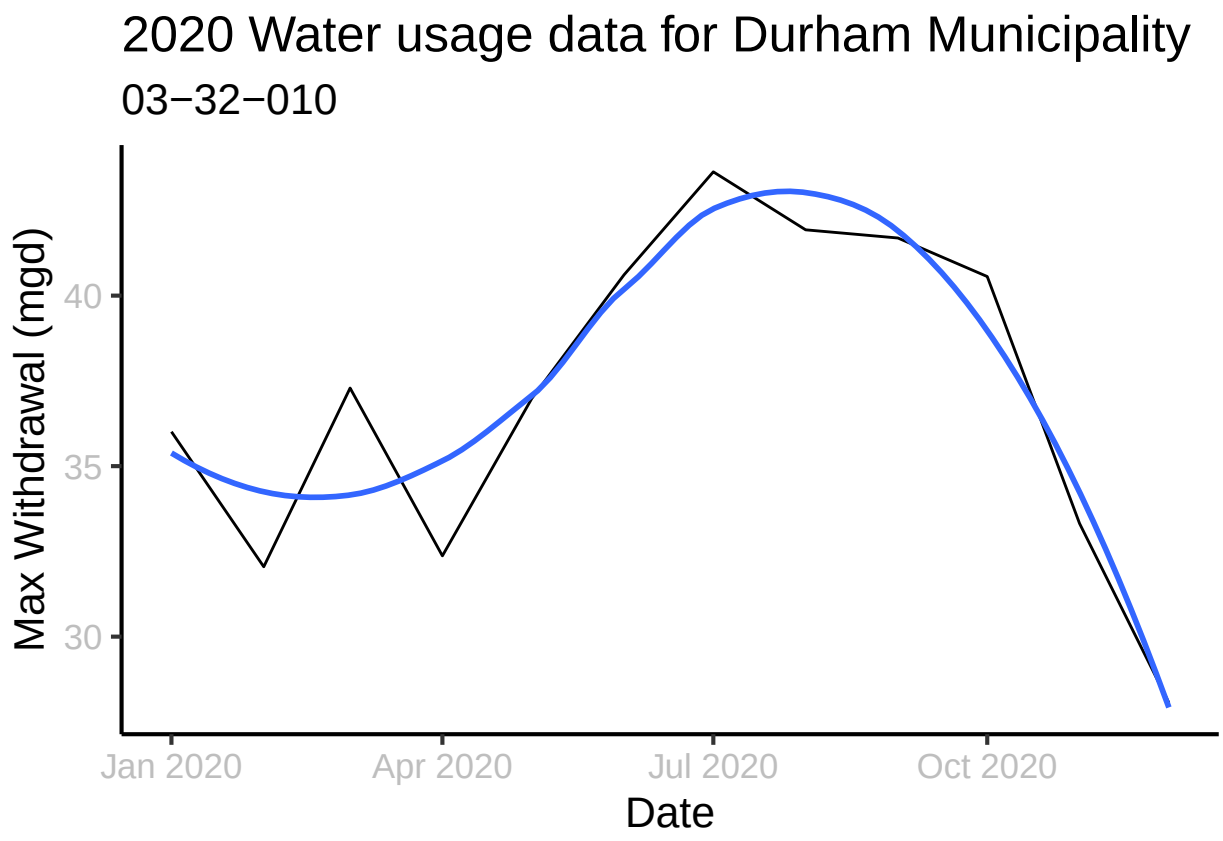
```
#4
DurhamH2O_2020 <- data.frame(
  "Month" = c('Jan','May','Sep','Feb','Jun','Oct',
              'Mar','Jul','Nov','Apr','Aug','Dec'),
  "Year" = rep(2020, 12),
  "WaterSystemName" = water.system.name,
  "PWSID" = pwsid,
  "Ownership" = ownership,
  "MaxDailyWthdrl" = as.numeric(max.withdrawals.mgd)) %>%
  mutate(Date = my(paste(Month, '-', Year))) %>%
  arrange(-desc(Date), .by_group = FALSE)

head(DurhamH2O_2020)
```

```
##   Month Year WaterSystemName     PWSID    Ownership MaxDailyWthdrl       Date
## 1   Jan 2020          Durham 03-32-010 Municipality          36.01 2020-01-01
## 2   Feb 2020          Durham 03-32-010 Municipality          32.05 2020-02-01
## 3   Mar 2020          Durham 03-32-010 Municipality          37.29 2020-03-01
## 4   Apr 2020          Durham 03-32-010 Municipality          32.37 2020-04-01
## 5   May 2020          Durham 03-32-010 Municipality          36.98 2020-05-01
## 6   Jun 2020          Durham 03-32-010 Municipality          40.61 2020-06-01
```

```
#5
A09_plot1 <- ggplot(DurhamH2O_2020, aes(x=Date, y=MaxDailyWthdrl)) +
  geom_line() +
  geom_smooth(method = 'loess', se = FALSE) +
  labs(title = paste("2020 Water usage data for",water.system.name,
                     ownership),
       subtitle = pwsid,
       y="Max Withdrawal (mgd)",
       x="Date")

print(A09_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 scraped**.

```
#6.
scraper <- function(thePWSID,theYear){
```

```
  theWebsite <- read_html(
    paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',
           thePWSID,'&year=',theYear))

  water.system.name <- theWebsite %>%
    html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>% html_text()
  pwsid <- theWebsite %>%
    html_nodes('td tr:nth-child(1) td:nth-child(5)') %>% html_text()
  ownership <- theWebsite %>%
    html_nodes('div+ table tr:nth-child(2) td:nth-child(4)') %>% html_text()
  max.withdrawals.mgd <- theWebsite %>%
    html_nodes('th~ td+ td') %>% html_text()

  df <- data.frame(
    "Month" = c('Jan','May','Sep','Feb','Jun','Oct',
                'Mar','Jul','Nov','Apr','Aug','Dec'),
    "Year" = rep(theYear, 12),
    "WaterSystemName" = water.system.name,
    "PWSID" = pwsid,
    "Ownership" = ownership,
    "MaxDailyWthdrl" = as.numeric(max.withdrawals.mgd)) %>%
      mutate(Date = my(paste(Month, '-', Year))) %>%
      arrange(-desc(Date), .by_group = FALSE)

  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
q7_df <- scraper('03-32-010','2015')

head(q7_df)
```

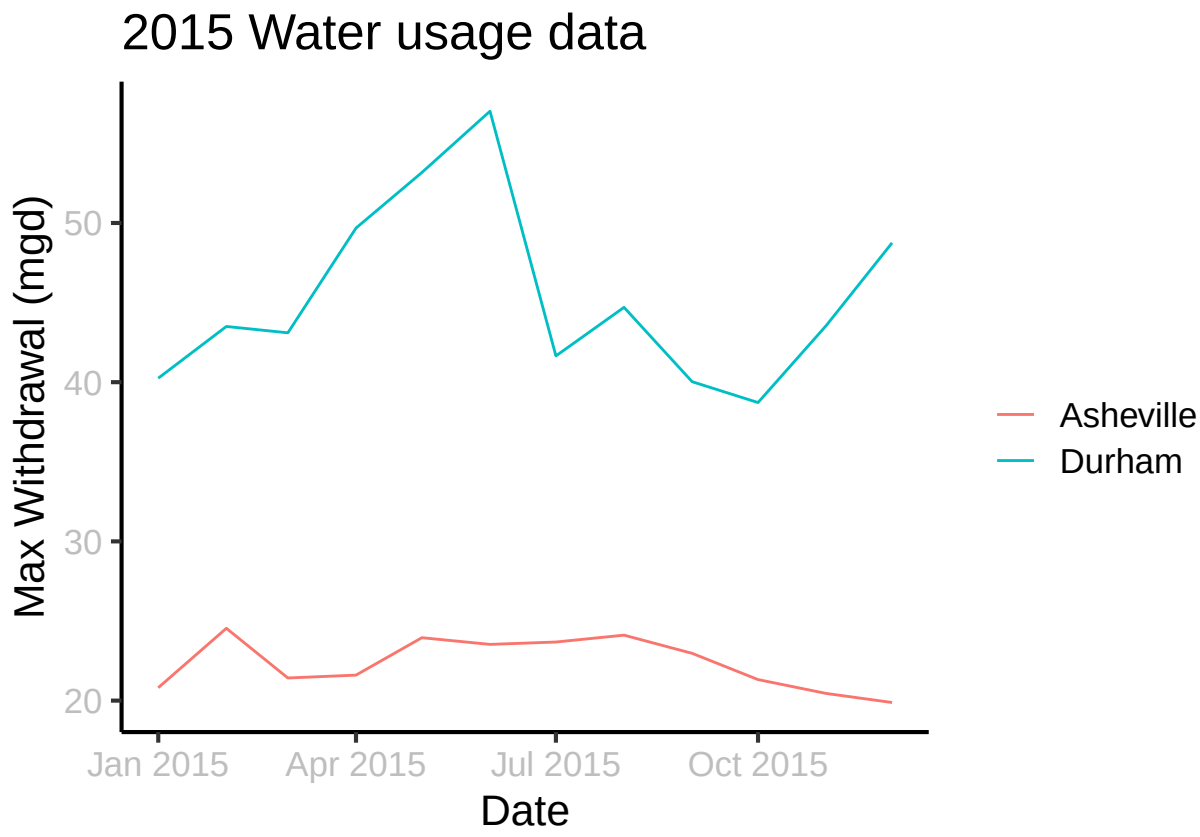
```
##   Month Year WaterSystemName     PWSID    Ownership MaxDailyWthdrl       Date
## 1   Jan 2015          Durham 03-32-010 Municipality          40.25 2015-01-01
## 2   Feb 2015          Durham 03-32-010 Municipality          43.50 2015-02-01
## 3   Mar 2015          Durham 03-32-010 Municipality          43.10 2015-03-01
## 4   Apr 2015          Durham 03-32-010 Municipality          49.68 2015-04-01
## 5   May 2015          Durham 03-32-010 Municipality          53.17 2015-05-01
## 6   Jun 2015          Durham 03-32-010 Municipality          57.02 2015-06-01
```

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 the Asheville to Durham’s water withdrawals.

```
#8
q8_df <- scraper('01-11-010','2015') %>%
  bind_rows(q7_df)

A09_plot2 <- ggplot(q8_df, aes(x=Date, y=MaxDailyWthdrl,
                               color = WaterSystemName)) +
  geom_line() +
  labs(title = '2015 Water usage data',
       y="Max Withdrawal (mgd)",
```

```
       x="Date",
       color='') +
  theme(legend.position = 'right')
print(A09_plot2)
```

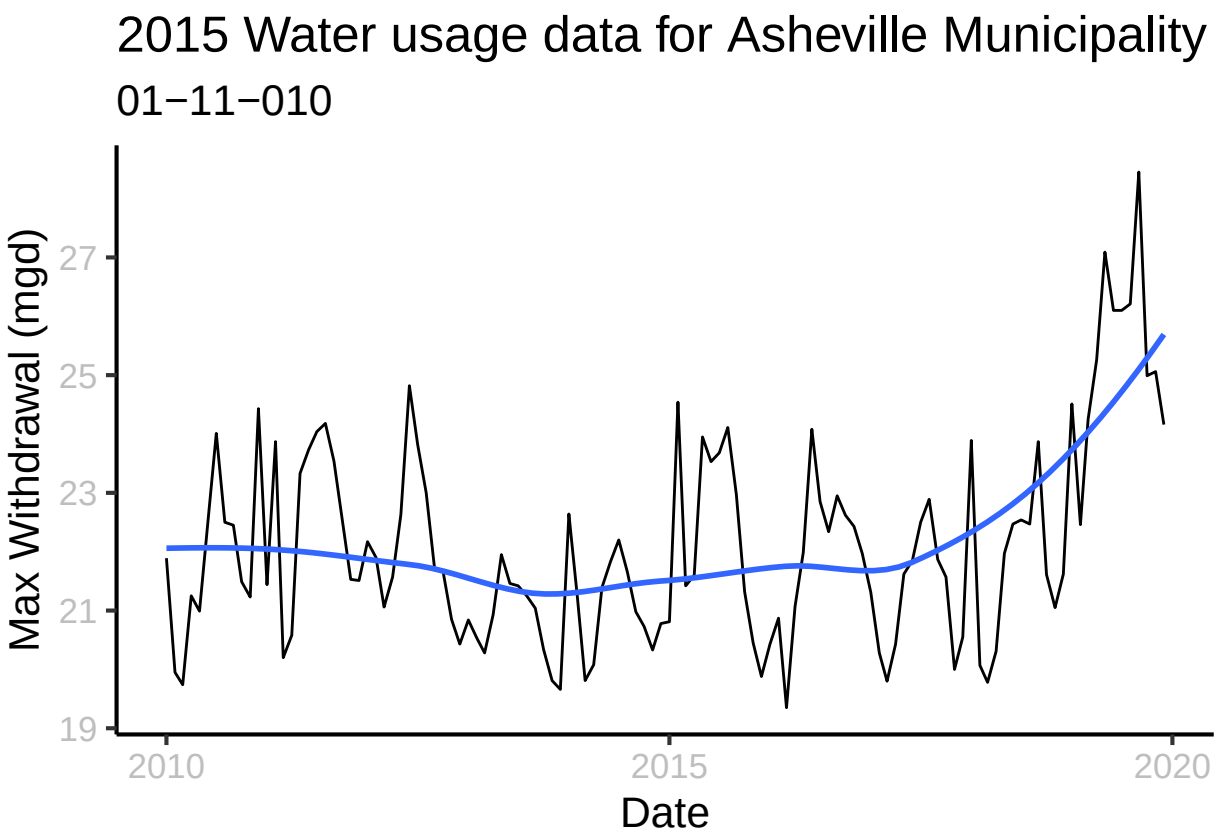


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

```
#9
q9_years <- as.character(c(2010:2019))

q9_df <- map2('01-11-010', q9_years, scraper) %>% bind_rows()

A09_plot3 <- ggplot(q9_df, aes(x=Date, y=MaxDailyWthdrl)) +
  geom_line() +
  geom_smooth(method = 'loess', se = FALSE) +
  labs(title = '2015 Water usage data for Asheville Municipality',
       subtitle = '01-11-010',
       y="Max Withdrawal (mgd)") +
  theme(legend.position = 'right')
print(A09_plot3)
```



Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time?

It does appear that Asheville has increased its water usage over the last 10 years.