Assignment 09: Data Scraping

Maggie O'Shea

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] "/Users/maggieoshea/Desktop/Spring 2022/Data Analytics/Environmental_Data_Analytics_2022/Assignm
```

```
library(tidyverse)
library(rvest)

A9theme <- theme_gray(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
            legend.position = "right",
            plot.title = element_text(size=14, hjust=0.5),
            plot.subtitle = element_text(size=10, hjust=0.5))

theme_set(A9theme)
```

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
LWSP_web <- 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 <- "div+ table tr:nth-child(1) td:nth-child(2)"
pswid <- "td tr:nth-child(1) td:nth-child(5)"
ownership <- "div+ table tr:nth-child(2) td:nth-child(4)"
max.withdrawals.mgd <- "th~ td+ td"
```

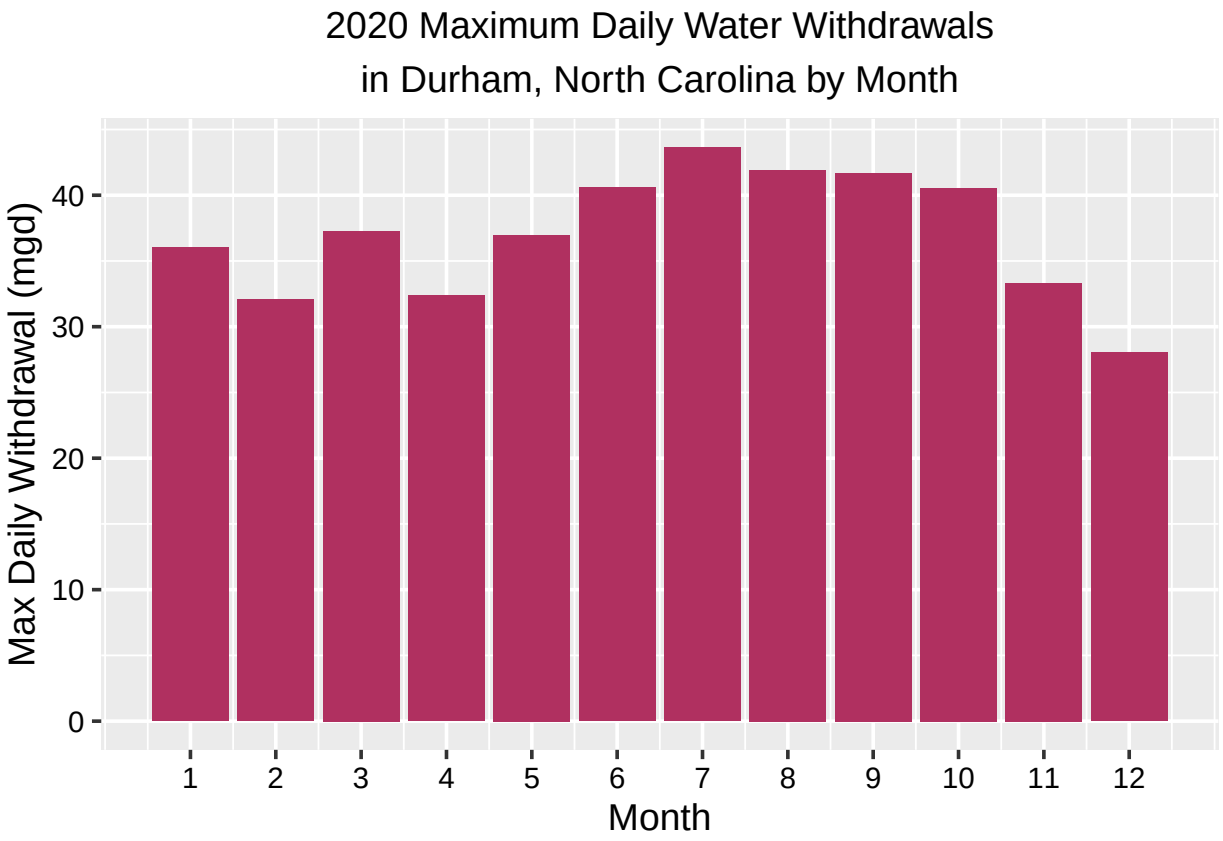
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   > 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

```r
#4
name_web <- LWSP_web %>% html_nodes(water.system.name) %>% html_text()
pswid_web <- LWSP_web %>% html_nodes(pswid) %>% html_text()
ownership_web <- LWSP_web %>% html_nodes(ownership) %>% html_text()
maxwithdrawals_web <- LWSP_web %>% html_nodes(max.withdrawals.mgd) %>% html_text()

months <- c("January", "May", "September", "February", "June", "October",
            "March", "July", "November", "April", "August", "December")
month.numbers <- c(1, 5, 9, 2, 6, 10, 3, 7, 11, 4, 8, 12)

df_Durhamwater2020 <- data.frame("WaterSystemName" = as.character(name_web),
                                 "PSWID" = as.character(pswid_web),
                                 "Ownership" = as.character(ownership_web),
                                 "MaxWithdrawals" = as.numeric(maxwithdrawals_web),
                                 "Month" = months,
                                 "MonthNumbers" = as.integer(month.numbers))



#5
ggplot(df_Durhamwater2020) +
  geom_col(aes(y = MaxWithdrawals, x = MonthNumbers), fill = "maroon")+
  labs(title = "2020 Maximum Daily Water Withdrawals",
       subtitle = "in Durham, North Carolina by Month",
       y = "Max Daily Withdrawal (mgd)",
       x = "Month")+
  scale_x_continuous(breaks = seq(1, 12, by = 1))+
  theme(plot.title = element_text(size=14, hjust=0.5),
        plot.subtitle = element_text(size=14, hjust=0.5))
```



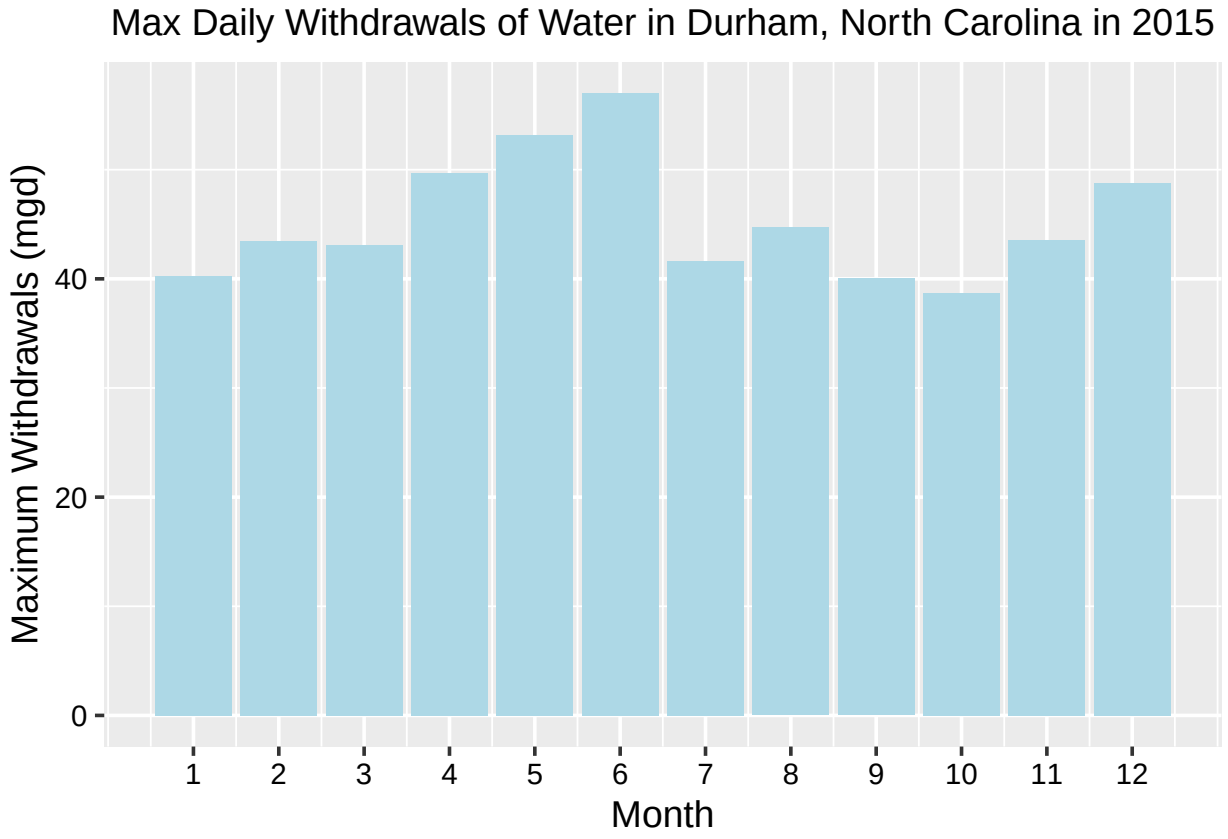
6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site scraped**.

```
#6.

water_scrape <- function(the_year, pwsid_code){

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

    water.system.name <- ".system-header"
    pwsid <- "td tr:nth-child(1) td:nth-child(5)"
    ownership <- "div+ table tr:nth-child(2) td:nth-child(4)"
    max.withdrawals.mgd <- "th~ td+ td"
    month <- ".fancy-table:nth-child(31) tr+ tr th"

    name_url <- the_url %>% html_nodes(water.system.name) %>% html_text()
    pwsid_url <- the_url %>% html_nodes(pwsid) %>% html_text()
    ownership_url <- the_url %>% html_nodes(ownership) %>% html_text()
    maxwithdrawals_url <- the_url %>% html_nodes(max.withdrawals.mgd) %>% html_text()

    months <- c("January", "May", "September", "February", "June", "October",
                "March", "July", "November", "April", "August", "December")

    dataframe <- data.frame("WaterSystemName" = name_url,
                                "PSWID" = pwsid_url,
```

```
                                  "Ownership" = ownership_url,
                                  "MaxWithdrawals" = as.numeric(maxwithdrawals_url),
                                  "Month" = months,
                                  "Year" = as.numeric(the_year),
                                  "MonthNumbers" = month.numbers)

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

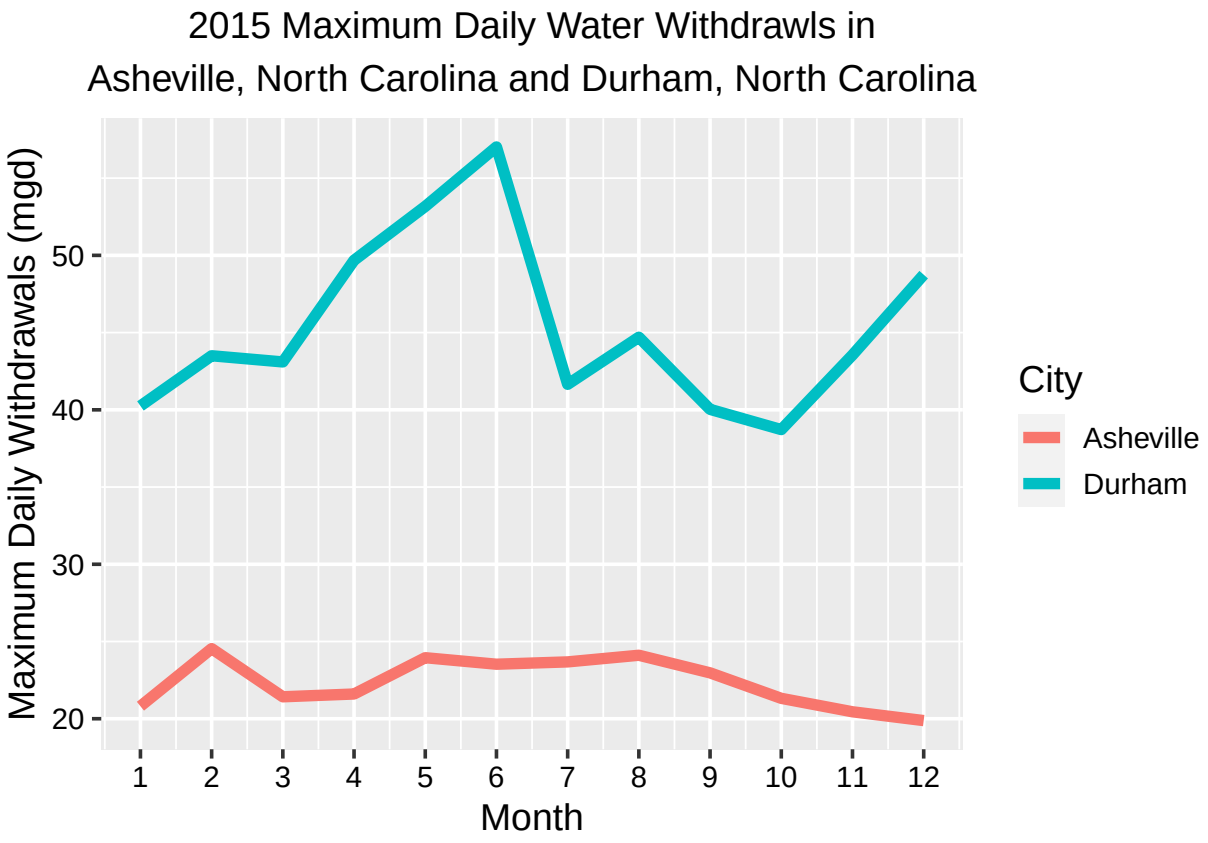
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID='03-32-010') for each month in 2015

```
#7

durham2015<- water_scrape(2015,"03-32-010")

ggplot(durham2015,aes(y = MaxWithdrawals, x=MonthNumbers)) +
  geom_col(fill = "lightblue")+
  labs(title = "Max Daily Withdrawals of Water in Durham, North Carolina in 2015",
       x = "Month",
       y = "Maximum Withdrawals (mgd)")+
  scale_x_continuous(breaks = seq(1, 12, by = 1))
```



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
asheville2015 <- water_scrape(2015, "01-11-010")



asheville.durham <- rbind(asheville2015, durham2015)


ggplot(asheville.durham, aes(x=MonthNumbers, y = MaxWithdrawals, color=WaterSystemName)) +
  geom_line(size = 2) +
  labs(title = "2015 Maximum Daily Water Withdrawls in",
       subtitle = "Asheville, North Carolina and Durham, North Carolina",
       x = "Month",
       y = "Maximum Daily Withdrawals (mgd)")+
       guides(color=guide_legend(title="City"))+
  theme(plot.title = element_text(size=14, hjust=0.5),
        plot.subtitle = element_text(size=14, hjust=0.5))+
  scale_x_continuous(breaks = seq(1, 12, by = 1))
```

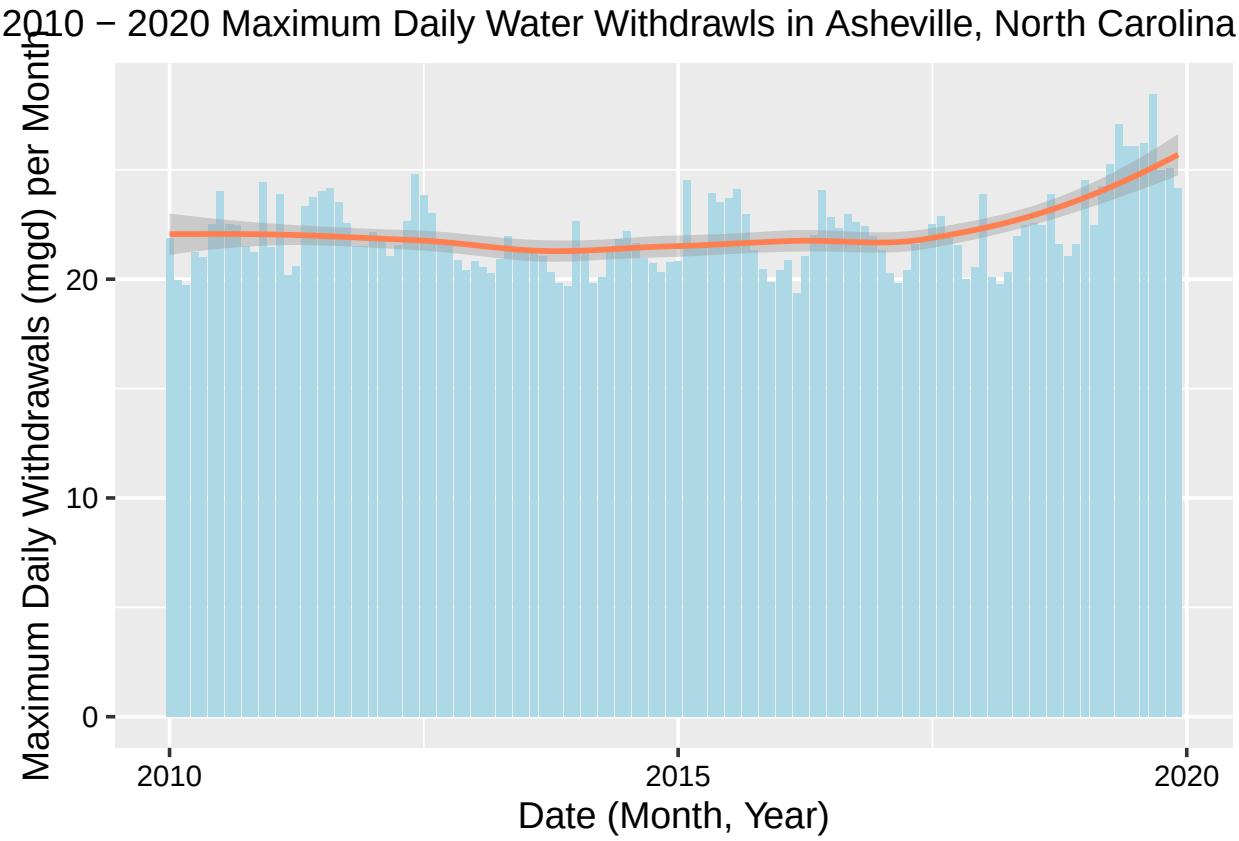


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

```
#9

years_asheville <- c(2010:2019)
```

```r
asheville_9years <- lapply(X =  years_asheville,
                  FUN = water_scrape,
                  pwsid_code="01-11-010")

asheville_9df <- bind_rows(asheville_9years)

asheville_final <- asheville_9df %>%
  mutate(MonthNumber = case_when(
    endsWith(Month, "anuary") ~ 1,
    endsWith(Month, "ebruary") ~ 2,
    endsWith(Month, "arch") ~ 3,
    endsWith(Month, "pril") ~ 4,
    endsWith(Month, "ay") ~ 5,
    endsWith(Month, "une") ~ 6,
    endsWith(Month, "uly") ~ 7,
    endsWith(Month, "ugust") ~ 8,
    endsWith(Month, "eptember") ~ 9,
    endsWith(Month, "ctober") ~ 10,
    endsWith(Month, "ovember") ~ 11,
    endsWith(Month, "ecember") ~ 12))

asheville_final$date <- paste( asheville_final$MonthNumber,"/1/", asheville_final$Year)

asheville_final$date <- gsub(" ", "", asheville_final$date)

asheville_final$date <- as.Date(asheville_final$date, "%m/%d/%Y")


ggplot(asheville_final , aes(x=date)) +
  geom_col(aes(y = MaxWithdrawals), fill="lightblue", linetype="twodash") +
  geom_smooth(aes(y= MaxWithdrawals), color = "coral", show.legend = FALSE)+
  labs(title = "2010 - 2020 Maximum Daily Water Withdrawls in Asheville, North Carolina, USA)",
       x = "Date (Month, Year)",
       y = "Maximum Daily Withdrawals (mgd) per Month")
```

```
## `geom_smooth()` using method = 'loess' and 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: Based on the plot, it appears that Asheville's water usage was steady until about 2017, after which time it began to increase.