Assignment 09: Data Scraping

Jess Ozog

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] "C:/ENV872/Environmental_Data_Analytics_2022/Assignments"
```

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

```
## Warning: package 'rvest' was built under R version 4.1.3
```

```
library(lubridate)

theme1 <- theme_bw(base_size=12) +
  theme(panel.border=element_rect(fill="transparent",color="gray",size=1),
        plot.title = element_text(hjust = 0.5), legend.position="right",
        plot.subtitle=element_text(hjust=0.5))
theme_set(theme1)
```

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham's 2020 Municipal Local Water Supply Plan (LWSP):

- Navigate to https://www.ncwater.org/WUDC/app/LWSP/search.php
- Change the date from 2021 to 2020 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
website <- read_html('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2020')
website
```

```
## {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
- PSWID
- Ownership
- From the "3. Water Supply Sources" section:
- Max 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 <- 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.withdrawals.mgd <- website %>%
  html_nodes("th~ td+ td , 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.)

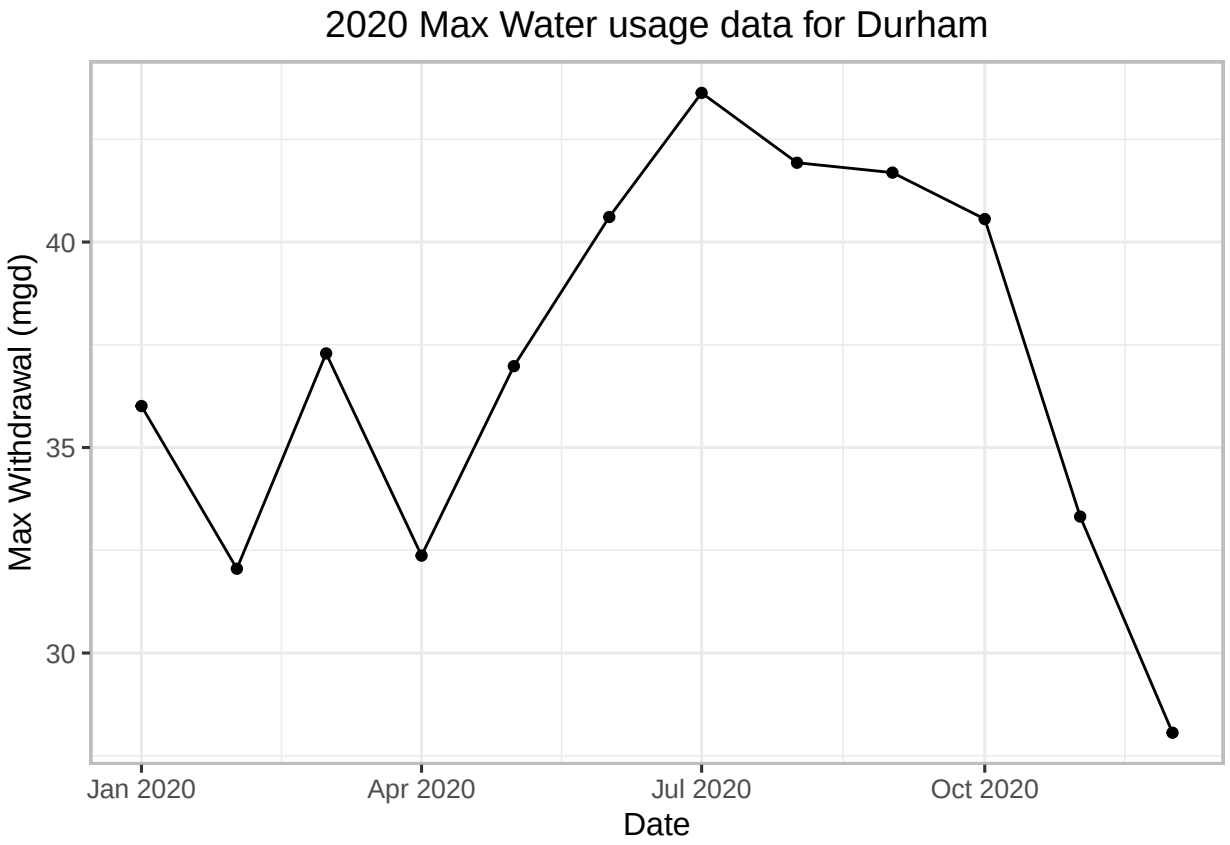
   TIP: Use `rep()` to repeat a value when creating a dataframe.

   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
withdrawl_df <- data.frame("Month" = c(1,5,9,2,6,10,3,7,11,4,8,12),
                           "Year" = rep(2020,12),
                           "System_Name" = water.system.name,
                           "PWSID" = pwsid,
                           "Ownership" = ownership,
                           "Max_Withdrawals_mgd" =
                             as.numeric(max.withdrawals.mgd)) %>%
  mutate(Date = my(paste(Month,"-",Year)))

#5
ggplot(withdrawl_df,aes(x=Date,y=Max_Withdrawals_mgd)) +
  geom_point() +
  geom_line() +
  labs(title = paste("2020 Max Water usage data for", water.system.name),
       y="Max Withdrawal (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 scraped**.

```
#6.
#Create our scraping function
scrape.it <- function(my_year, my_pwsid){

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

  the_water.system <- 'div+ table tr:nth-child(1) td:nth-child(2)'
  the_pwsid <- 'td tr:nth-child(1) td:nth-child(5)'
  the_ownership <- 'div+ table tr:nth-child(2) td:nth-child(4)'
  the_max.withdrawals <- 'th~ td+ td , th~ td+ td'

  the_water.system <- the_website %>% html_nodes(the_water.system) %>%
        html_text()
  the_pwsid <- the_website %>%   html_nodes(the_pwsid) %>%
        html_text()
  the_ownership <- the_website %>% html_nodes(the_ownership) %>%
        html_text()
  the_max.withdrawals <- the_website %>% html_nodes(the_max.withdrawals) %>%
        html_text()
```

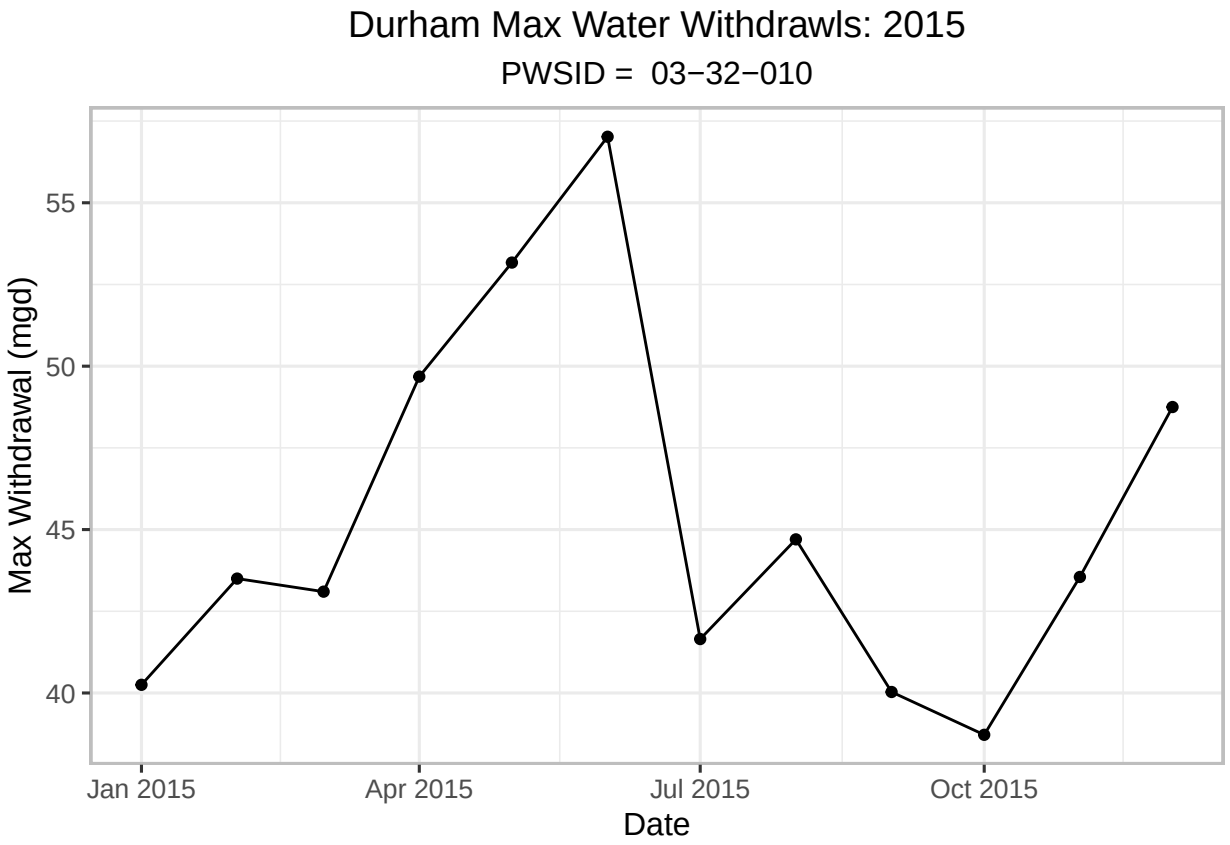
```
  withdrawls_df2 <- data.frame("Month" = c(1,5,9,2,6,10,3,7,11,4,8,12),
                               "Year" = rep(my_year,12),
                               "Max_Withdrawals_mgd" =
                                 as.numeric(the_max.withdrawals)) %>%
    mutate(Water_system_name = !!the_water.system,
           PWSID = !!the_pwsid,
           Ownership = !!the_ownership,
           Date = my(paste(Month,"-",Year)))

  return(withdrawls_df2)
}
```

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 <- scrape.it(2015,'03-32-010')
view(durham2015)


ggplot(durham2015,aes(x=Date,y=Max_Withdrawals_mgd)) +
  geom_point() +
  geom_line() +
  labs(title = paste("Durham Max Water Withdrawls: 2015"),
       subtitle = paste("PWSID = ", pwsid),
       y="Max Withdrawal (mgd)",
       x="Date")
```

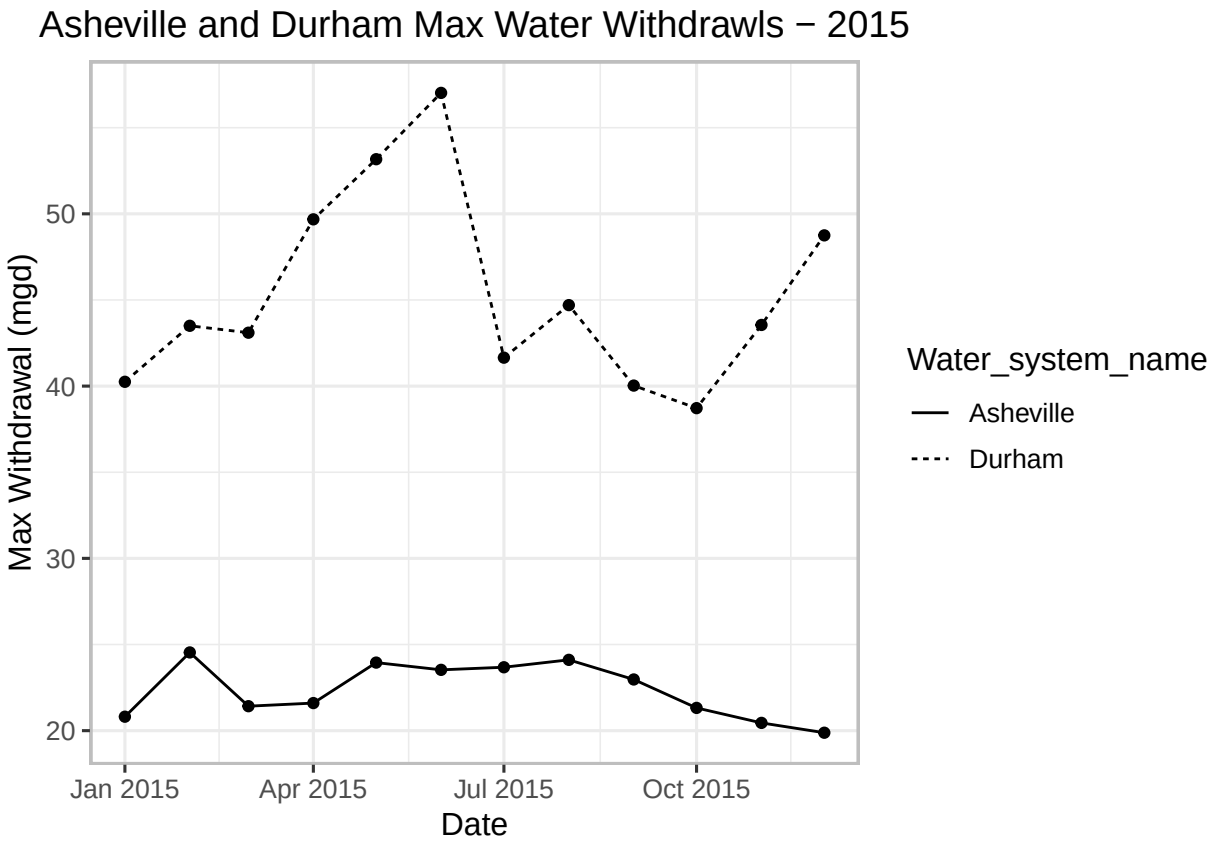


8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares the Asheville to Durham's water withdrawals.

```
#8
asheville2015 <-scrape.it(2015,'01-11-010')
view(asheville2015)

AshDurCombo <- rbind(durham2015, asheville2015)

ggplot(AshDurCombo, aes(x=Date, y=Max_Withdrawals_mgd)) +
  geom_line(aes(linetype=Water_system_name)) +
  geom_point() +
  labs(title = "Asheville and Durham Max Water Withdrawls - 2015",
       x="Date",
       y="Max Withdrawal (mgd)")
```



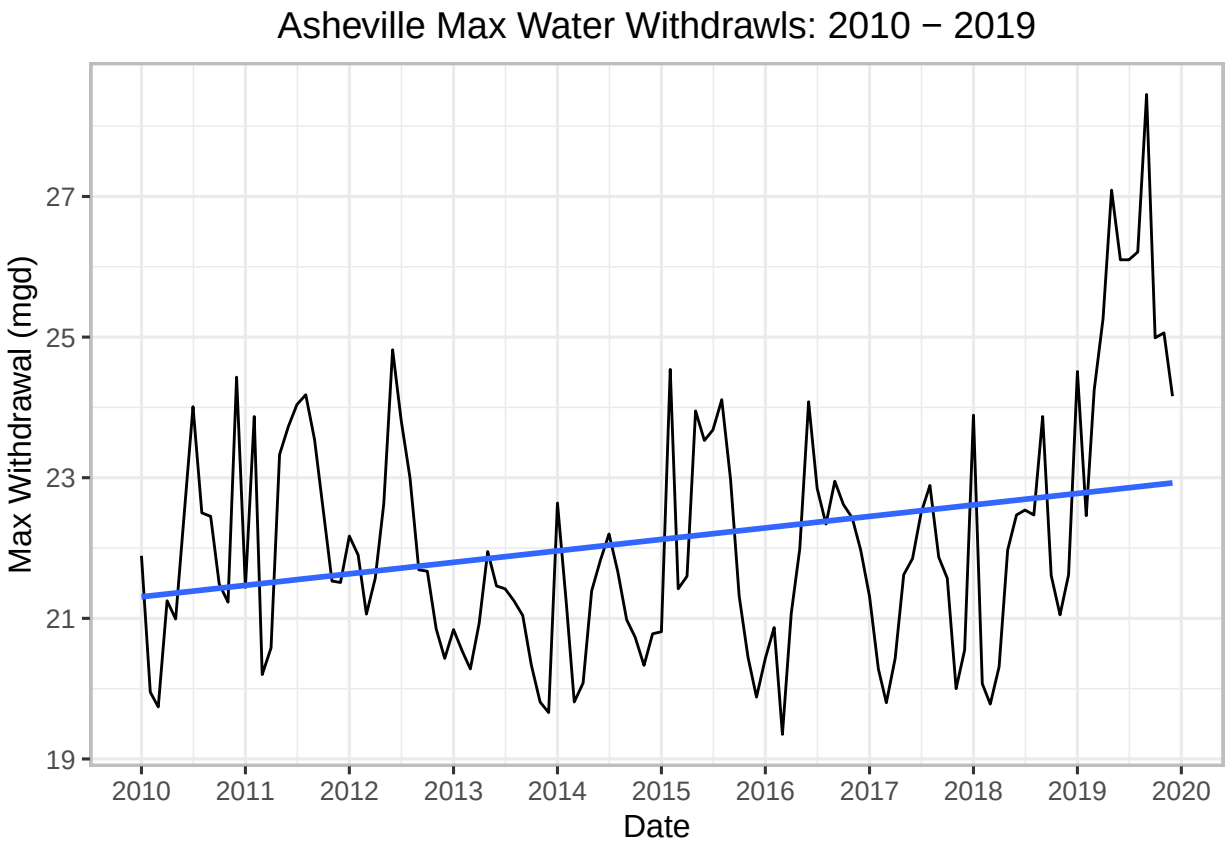
9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

```
#9
asheville2010 <- scrape.it(2010,'01-11-010')
asheville2011 <- scrape.it(2011,'01-11-010')
asheville2012 <- scrape.it(2012,'01-11-010')
asheville2013 <- scrape.it(2013,'01-11-010')
asheville2014 <- scrape.it(2014,'01-11-010')
asheville2016 <- scrape.it(2016,'01-11-010')
asheville2017 <- scrape.it(2017,'01-11-010')
asheville2018 <- scrape.it(2018,'01-11-010')
asheville2019 <- scrape.it(2019,'01-11-010')

asheville2010_2019 <- rbind(asheville2010,asheville2011,asheville2012,
                      asheville2013,asheville2014,asheville2015,asheville2016,
                      asheville2017,asheville2018,asheville2019)

ggplot(asheville2010_2019, aes(x=Date, y=Max_Withdrawals_mgd)) +
  geom_line() +
  geom_smooth(method="lm", se=FALSE) +
  scale_x_date(date_labels="%Y", date_breaks="1 year") +
  labs(title = "Asheville Max Water Withdrawls: 2010 - 2019",
       x="Date",
       y="Max 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? Yes, Asheville has a trend in water usage over time. From 2010 to 2019, the maximum daily withdrawl of water has gradually increased.