Assignment 09: Data Scraping

Austin Guimond

OVERVIEW

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

Directions

1. Rename this file `<FirstLast>_A09_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 to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.

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] "/home/guest/R/EDA-Fall2022"
```

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

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

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham's 2021 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=2021

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

```
#2
#Website to be scraped
webpage <- read_html('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2021')
webpage
```

```
## {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:
- Maximum Daily 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), with the first value being "27.6400".

```
#3
water.system.name <- webpage %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()
water.system.name
```

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

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

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

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

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

```
max.withdrawals.mgd <-  webpage %>%
  html_nodes("th~ td+ td") %>%
  html_text()
max.withdrawals.mgd
```

```
##  [1] "27.6400" "41.7900" "36.7200" "27.9700" "37.9500" "42.2400" "30.5400"
##  [8] "43.6200" "31.2800" "33.7600" "46.0800" "29.7800"
```

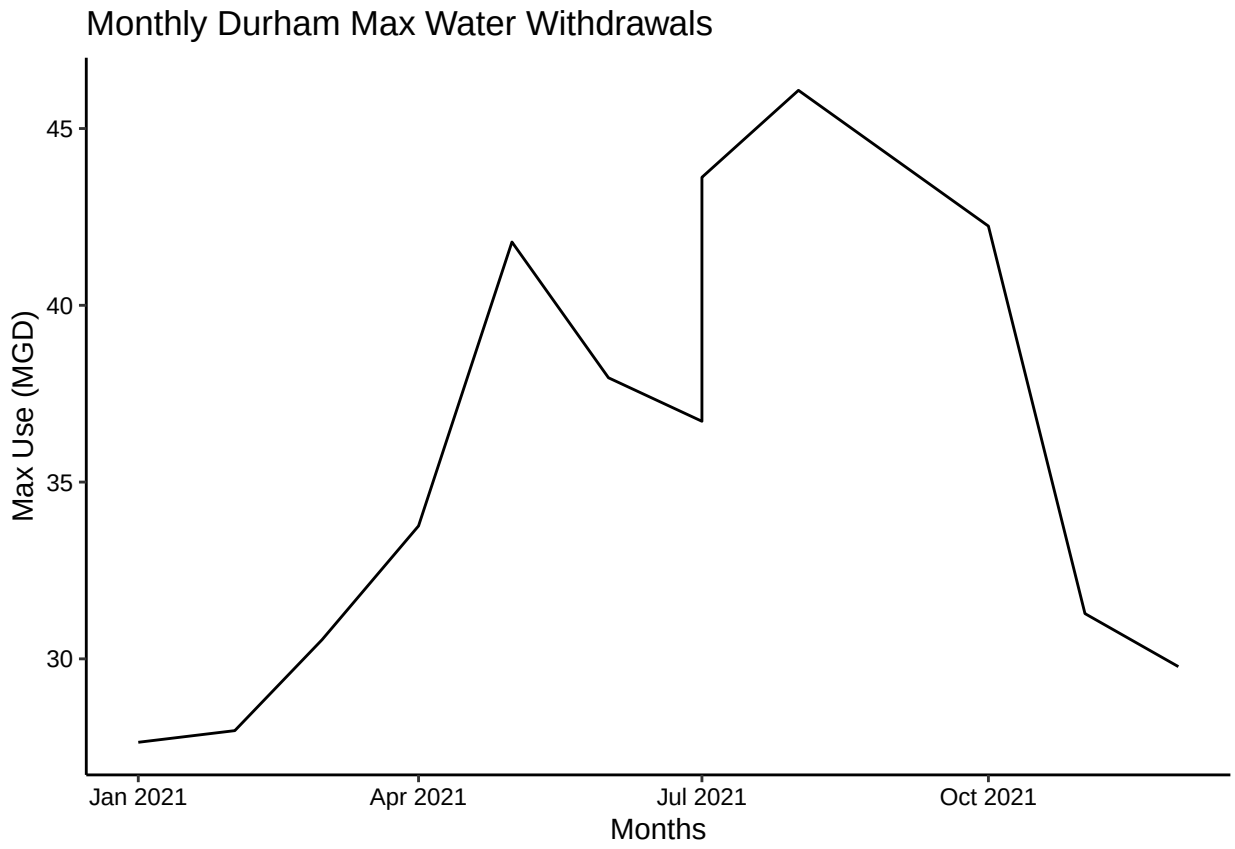
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   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...

5. Create a line plot of the maximum daily withdrawals across the months for 2021

```
#4
#Create MGD dataframe
Withdrawals_df <- data.frame("Month" = c(1,5,7,2,6,10,3,7,11,4,8,12),
                             "Year" = rep(2021,12),
                             "Max_Withdrawals.mgd" = as.numeric(max.withdrawals.mgd )) %>%
  mutate(Water_System_Name = !!water.system.name,
         Ownership = !!ownership,
         PSWID = !!pswid,
         Date = my(paste(Month,"-",Year)))
#5
#Plot 2021 MGD
Max.Withdrawals.Plot <- ggplot(Withdrawals_df, aes(x = Date, y = Max_Withdrawals.mgd)) +
  geom_line()+
 ggtitle("Monthly Durham Max Water Withdrawals")+
  xlab("Months") + ylab("Max Use (MGD)")
print(Max.Withdrawals.Plot)
```



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.
#create base url
the_base_url <- 'https://www.ncwater.org/WUDC/app/LWSP/report.php'
PWSID <- '03-32-010'
the_year <- '2018'
scrape_url <- paste0( the_base_url, '?pwsid=', PWSID, '&year=', the_year)
print(scrape_url)
```

```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2018"
```

```
#retrieve website contents
the_website <- read_html(scrape_url)

#find scraped values
water.system.name <- the_website %>% html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>% html_
water.system.name
```

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

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

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

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

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

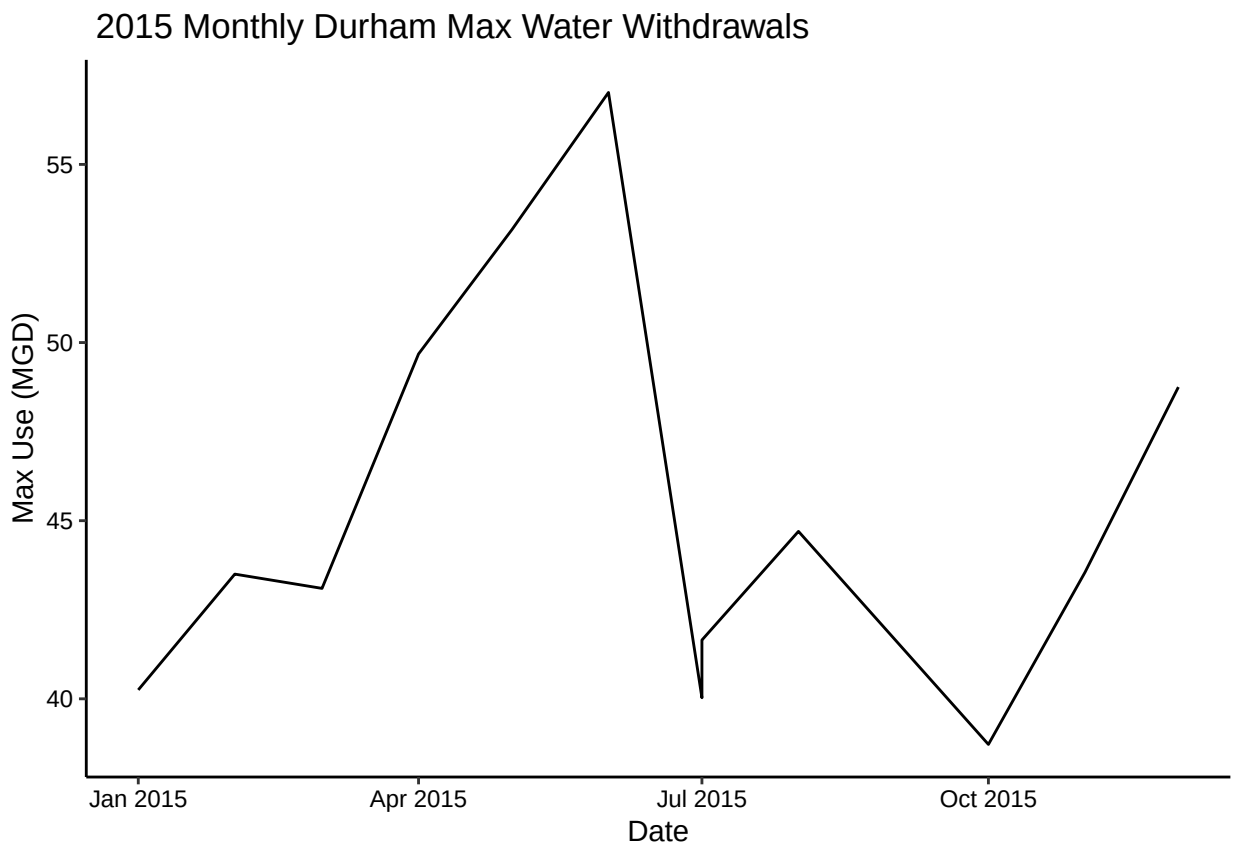
```
max.withdrawals.mgd <-  the_website %>% html_nodes("th~ td+ td") %>% html_text()
max.withdrawals.mgd
```

```
##  [1] "34.0700" "32.6800" "44.1500" "36.8300" "30.1300" "33.0300" "33.4100"
##  [8] "45.1500" "30.6100" "33.5300" "42.5400" "30.8400"
```

```
#create scrapping function
scrape.it <- function(the_year, PWSID) {

Scrape_website <- read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php', '?pwsid=', PWSID

#find scraped values
  Water_System_tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'
  pswid_tag <- 'td tr:nth-child(1) td:nth-child(5)'
  owner_tag <- 'div+ table tr:nth-child(2) td:nth-child(4)'
  max_withdrawals_tag <- 'th~ td+ td'

water.system.name <- Scrape_website %>% html_nodes(Water_System_tag) %>% html_text()
pswid <- Scrape_website %>% html_nodes(pswid_tag) %>% html_text()
ownership <- Scrape_website %>% html_nodes(owner_tag) %>% html_text()
max.withdrawals.mgd <-  Scrape_website %>% html_nodes(max_withdrawals_tag) %>% html_text()

#create data frame
Withdrawals_Frame <- data.frame("Month" = c(1,5,7,2,6,10,3,7,11,4,8,12),
                                "Year" = rep(the_year,12),
                              "Max_Withdrawals.mgd" = as.numeric(max.withdrawals.mgd )) %>%
  mutate(Water_System_Name = !!water.system.name,
         Ownership = !!ownership,
         PWSID = !!pswid,
          Date = my(paste(Month,"-",Year)))
#return dataframe
 return(Withdrawals_Frame)}

# test function
dat_frame.2018 <- scrape.it(2018, '03-32-010' )
```

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

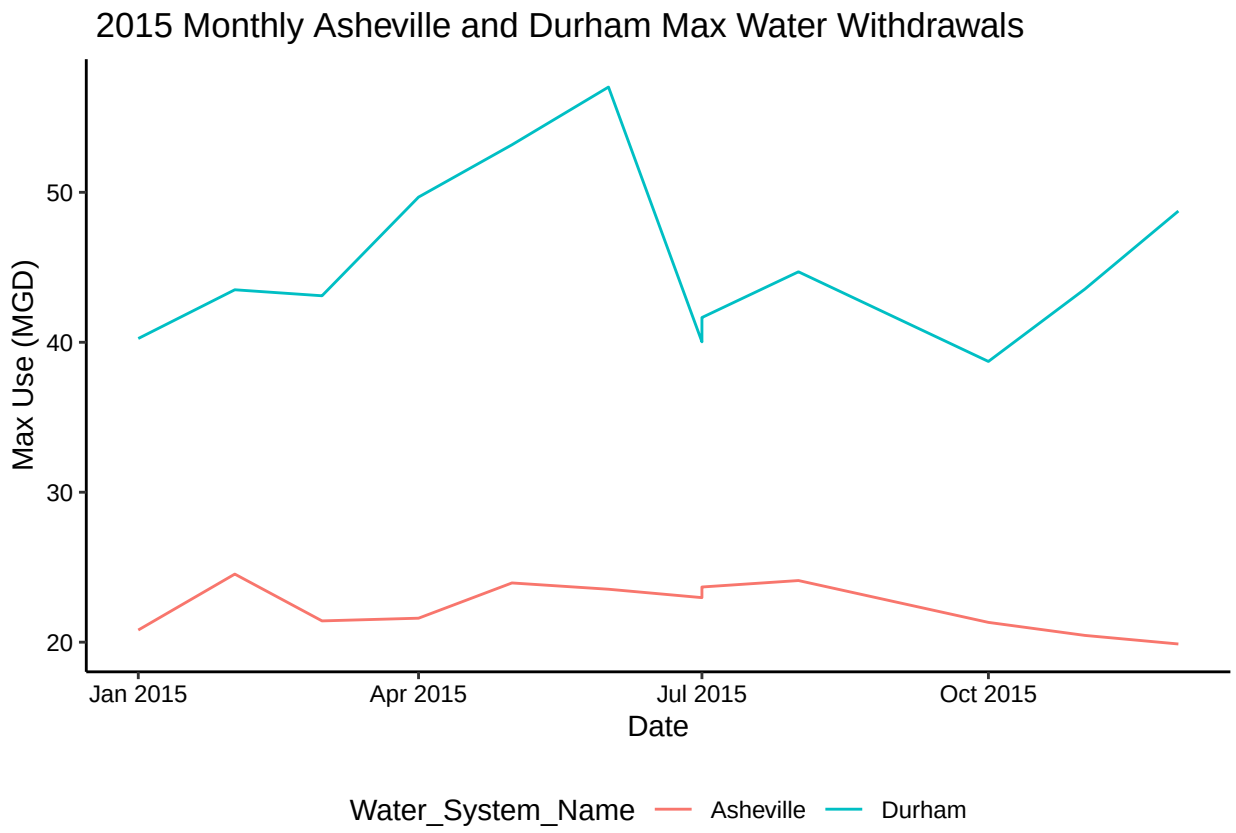
```
#7
Durham_frame.2015 <- scrape.it(2015, '03-32-010')


Max.Withdrawals.2015.Plot <- ggplot(Durham_frame.2015, aes(x = Date, y = Max_Withdrawals.mgd)) +
  geom_line()+
 ggtitle(" 2015 Monthly Durham Max Water Withdrawals")+
  xlab("Date") + ylab("Max Use (MGD)")
print(Max.Withdrawals.2015.Plot)
```



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
#Scrape Asheville 2015 data
Asheville_frame.2015 <- scrape.it(2015, '01-11-010')

#Combine 2015 Ashville and Durham Data sets
Full_data_2015 <- rbind(Asheville_frame.2015, Durham_frame.2015)

#Plot 2015 MGD for Durham and Ashville
Durham_Asheville_2015_Plot <- ggplot(Full_data_2015, aes(x = Date, y = Max_Withdrawals.mgd, color= Wate
  geom_line()+
 ggtitle(" 2015 Monthly Asheville and Durham Max Water Withdrawals")+
```

```
  xlab("Date") + ylab("Max Use (MGD)")
print(Durham_Asheville_2015_Plot)
```

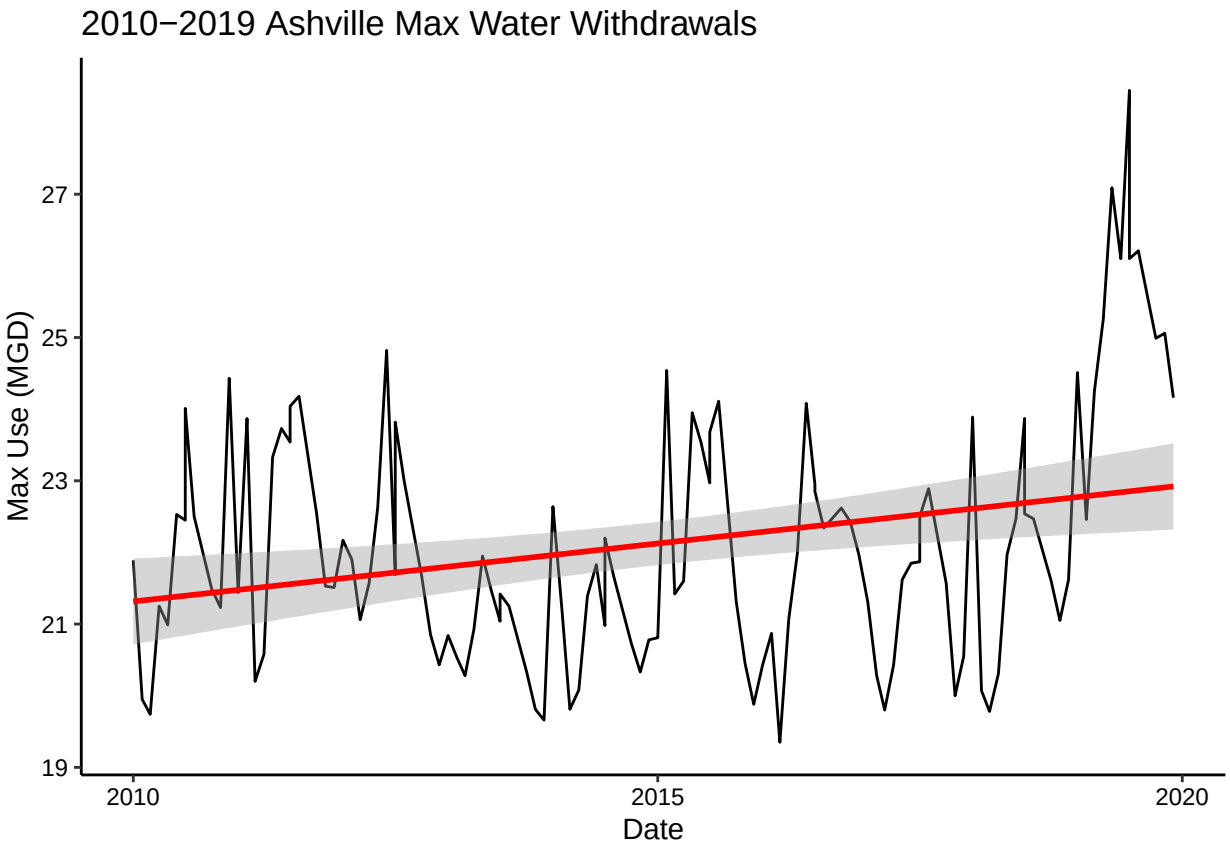


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

   TIP: See Section 3.2 in the "09_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
#Pull all Ashville data from 2010-2019
Asheville_2010_2019 <- map2(2010:2019, '01-11-010', scrape.it)

#create full data frame
Asheville_DF <- bind_rows(Asheville_2010_2019)

Asheville_Ten_Year_Plot <- ggplot(Asheville_DF, aes(x = Date, y = Max_Withdrawals.mgd)) +
  geom_line()+
  geom_smooth(method=lm, col= 'red')+
 ggtitle("2010-2019 Ashville Max Water Withdrawals")+
  xlab("Date") + ylab("Max Use (MGD)")
print(Asheville_Ten_Year_Plot)
```

```
## `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? There is a trend of increasing water use in Ashville between 2010 and 2019. There is an especially large jump in usage in 2019. It is possible that Ashville's population has increased over the ten years resulting in increased water use.