Assignment 09: Data Scraping

Aurora McCollum

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 Use max daily use!! not average!

```
#1
getwd()
```

```
## [1] "/Users/rorymccollum/Desktop/Rdata/Environmental_Data_Analytics_2022/Assignments"
```

```
library(tidyverse)
library(lubridate)
library(viridis)
library(rvest)
library(dataRetrieval)

rorystheme<-theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        panel.background = element_rect(fill = "blanchedalmond"),
        legend.position = "right")

theme_set(rorystheme)
```

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 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
webpage <- read_html('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2020')
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:
- 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 <-  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] "36.0100" "36.9800" "41.6900" "32.0500" "40.6100" "40.5600" "37.2900"
##  [8] "43.6300" "33.3200" "32.3700" "41.9300" "28.0600"
```

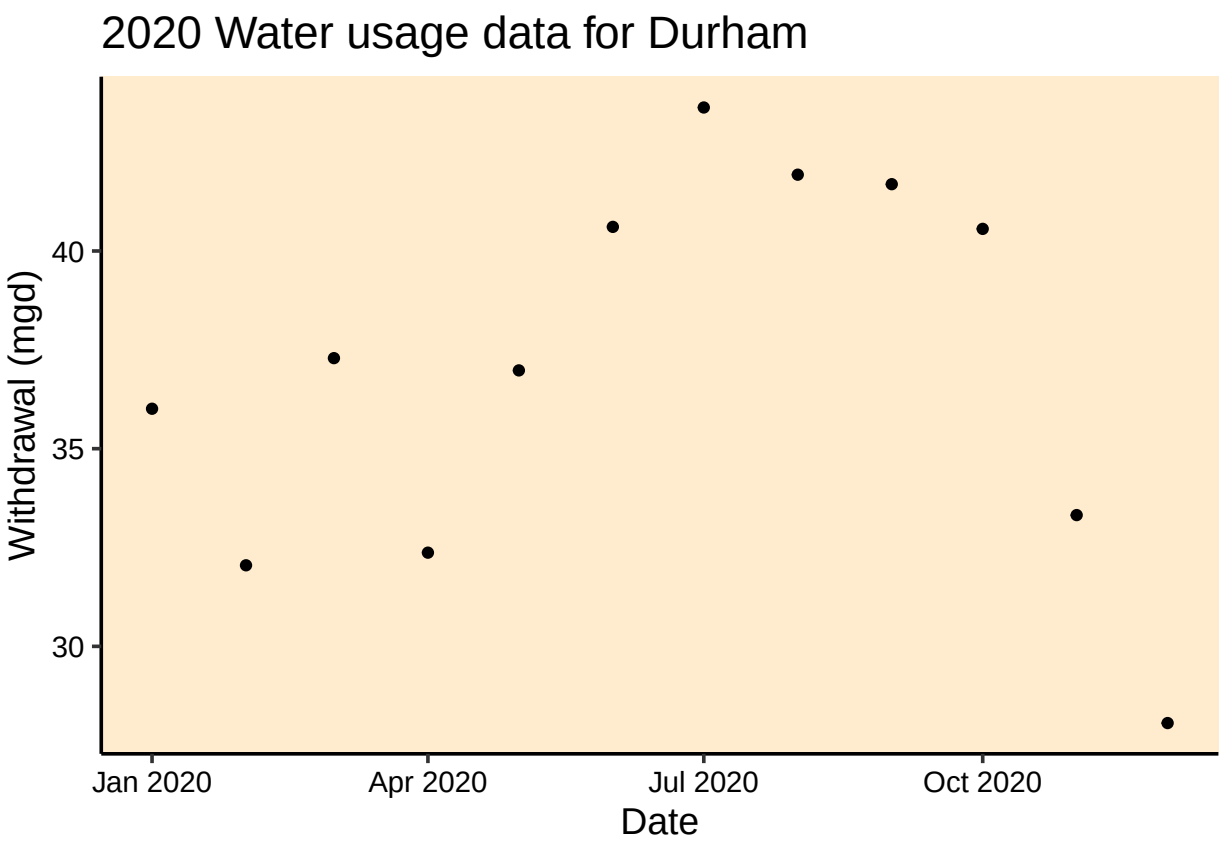
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   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
weirdmonths<-c("Jan", "May", "Sept", "Feb", "Jun", "Oct", "Mar", "Jul", "Nov", "Apr", "Aug", "Dec")

#date1<-make_date(year = 2020, month = )

df_watersystem <- data.frame("Month" = (weirdmonths),
                             "Year" = rep(2020,12),
                             "Max_Withdrawals_mgd" = as.numeric(max.withdrawals.mgd),
                             "Ownership" = (ownership),
                             "PWSID"=(pswid),
                             "Water_System_Name"= (water.system.name)) %>%
  mutate(Date=my(paste(Month,"-",Year)))



#5

ggplot(df_watersystem,aes(x=Date,y=Max_Withdrawals_mgd)) +
  geom_point() +
  labs(title = paste("2020 Water usage data for",water.system.name),
       y="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**.

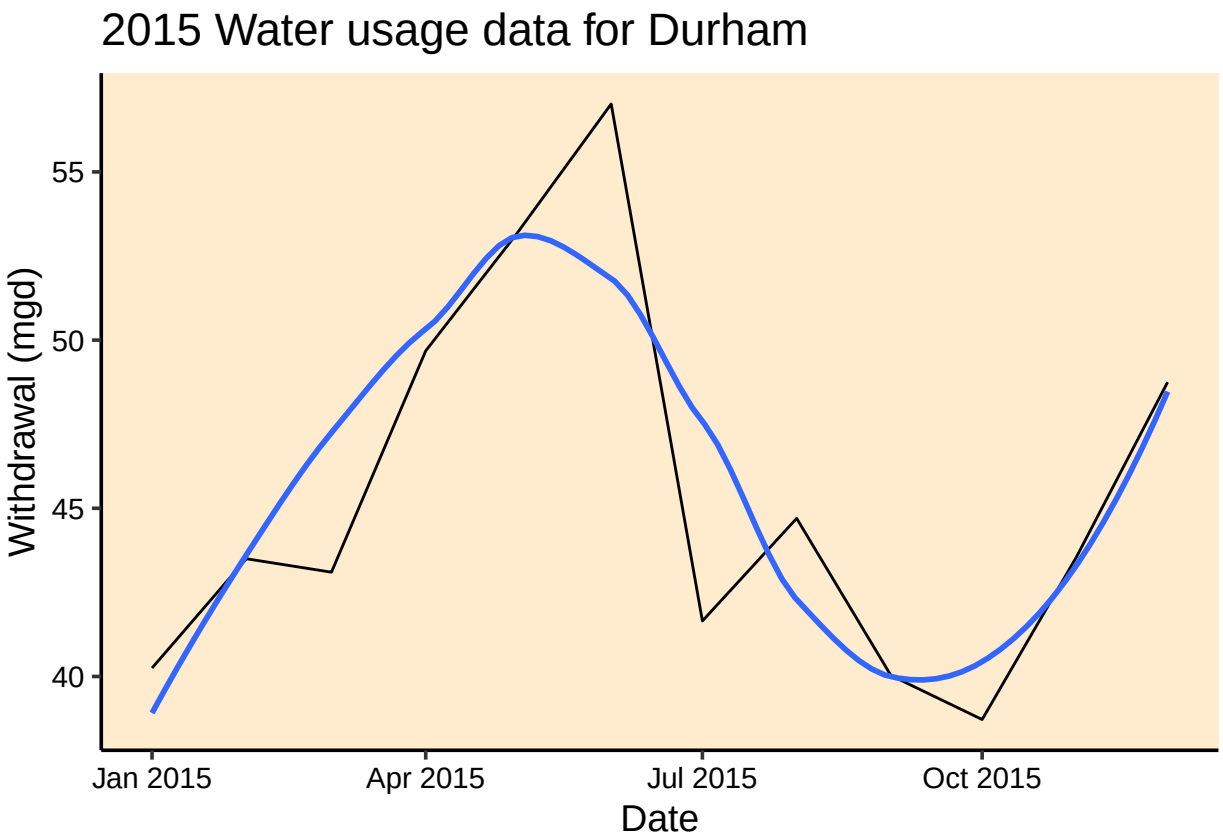
```
#6.

scrapeit<-function(the_year, the_pswid2){

  #retrieve website
The_website <- read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',the_pswid2,'&

water.system.name2 <-  The_website %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()

pswid2 <-  The_website %>% html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
  html_text()

ownership2 <-  The_website %>% html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%
  html_text()

max.withdrawals.mgd2 <-  The_website %>% html_nodes("th~ td+ td") %>%
  html_text()

weirdmonths2<-c("Jan", "May", "Sept", "Feb", "Jun", "Oct", "Mar", "Jul", "Nov", "Apr", "Aug", "Dec")

df_watersystem2 <- data.frame("Month" = (weirdmonths2),
                              "Year" = rep(the_year,12),
```

```
                              "Max_Withdrawals_mgd" = as.numeric(max.withdrawals.mgd2),
                              "Ownership" = (ownership2),
                              "PWSID"=(pswid2),
                              "Water_System_Name"= (water.system.name2)) %>%
  mutate(Date=my(paste(Month,"-",Year)))

return(df_watersystem2)
}
```

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

```
#7
withdrawl2015<-scrapeit(2015,'03-32-010')

ggplot(withdrawl2015,aes(x=Date,y=Max_Withdrawals_mgd)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2015 Water usage data for",water.system.name),
       y="Withdrawal (mgd)",
       x="Date")
```

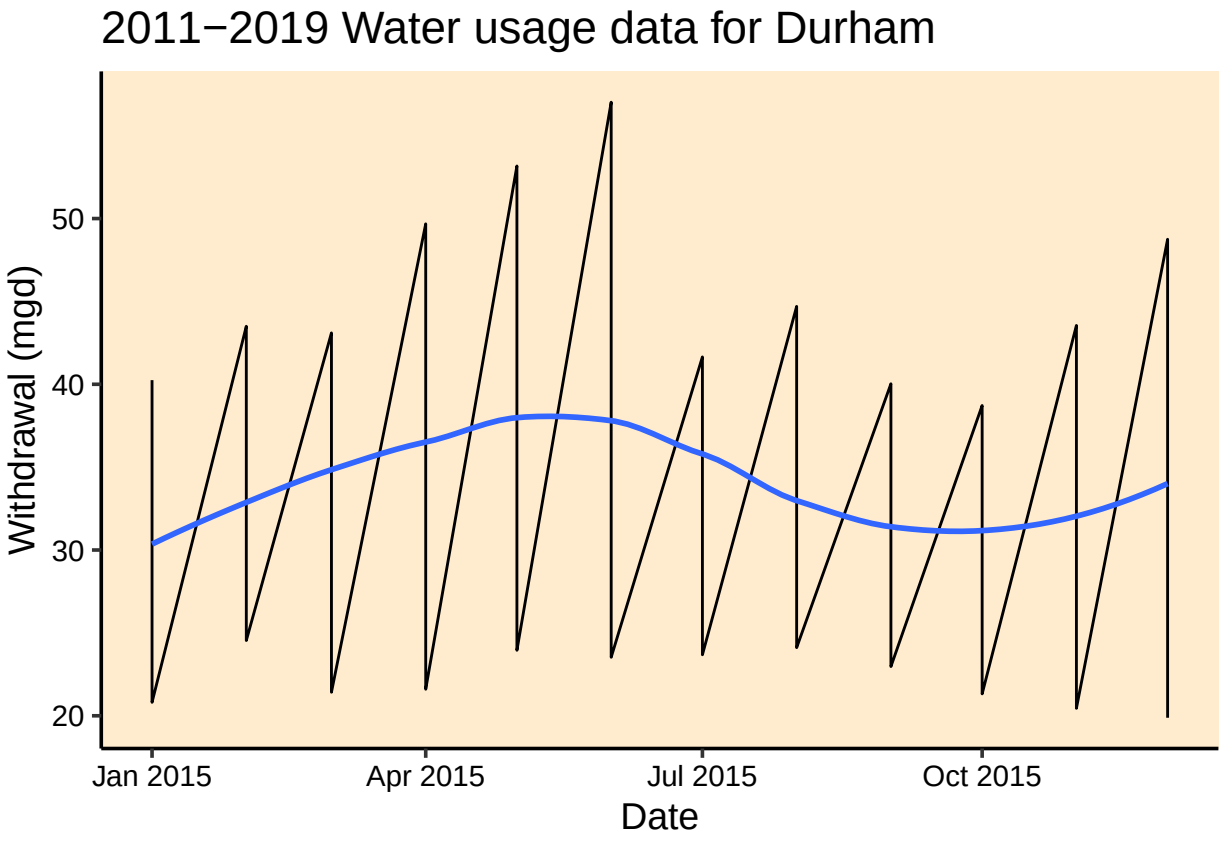
```
## `geom_smooth()` using formula 'y ~ x'
```



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

withdrawl2015.Ash<-scrapeit(2015,'01-11-010')

df_ash.durh<-bind_rows(withdrawl2015,withdrawl2015.Ash)

ggplot(df_ash.durh,aes(x=Date,y=Max_Withdrawals_mgd)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2011-2019 Water usage data for",water.system.name),
       y="Withdrawal (mgd)",
       x="Date")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

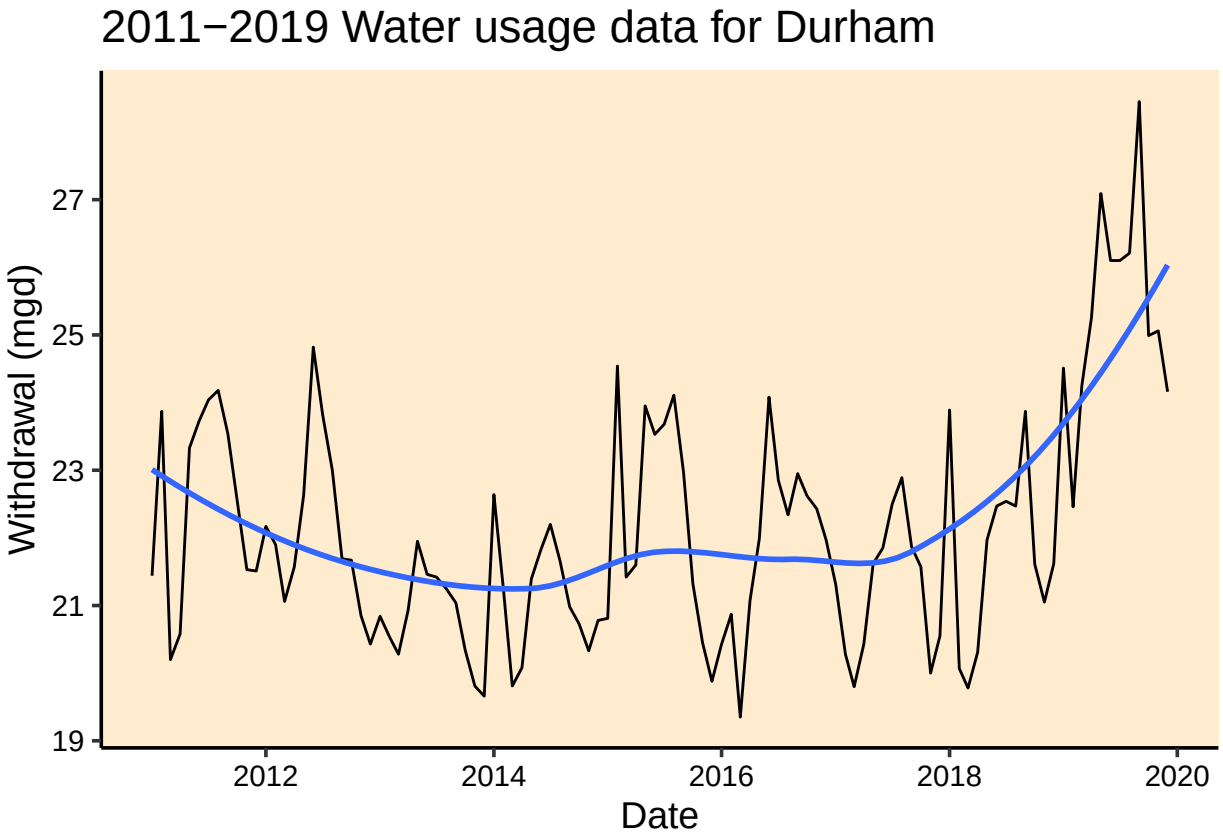


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

```
#9
the_years<-seq(2011,2019)
the_facility<-'01-11-010'

withdrawal_10.19<- lapply(X=the_years,
                          FUN= scrapeit,
                          the_facility)
df_final<-bind_rows(withdrawal_10.19)

ggplot(df_final,aes(x=Date,y=Max_Withdrawals_mgd)) +
```

```
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2011-2019 Water usage data for",water.system.name),
       y="Withdrawal (mgd)",
       x="Date")
```

```
## `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? It looks like water usage has increased over all since 2011. There is a trend, potentially an exponential one.