Assignment 09: Data Scraping

Rachel Gordon

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() #check working directory
```

```
## [1] "/Users/rachelgordon/Documents/Duke Spring 2022/Environmental Data Analytics/Environmental_Data_
```

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

#building my theme
RGtheme <- theme_gray(base_size = 12) +
  theme(axis.text = element_text(color = "black"))+
  theme(plot.title = element_text(hjust = 0.5))
```

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

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

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

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

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

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

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

```
#average daily use
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.)

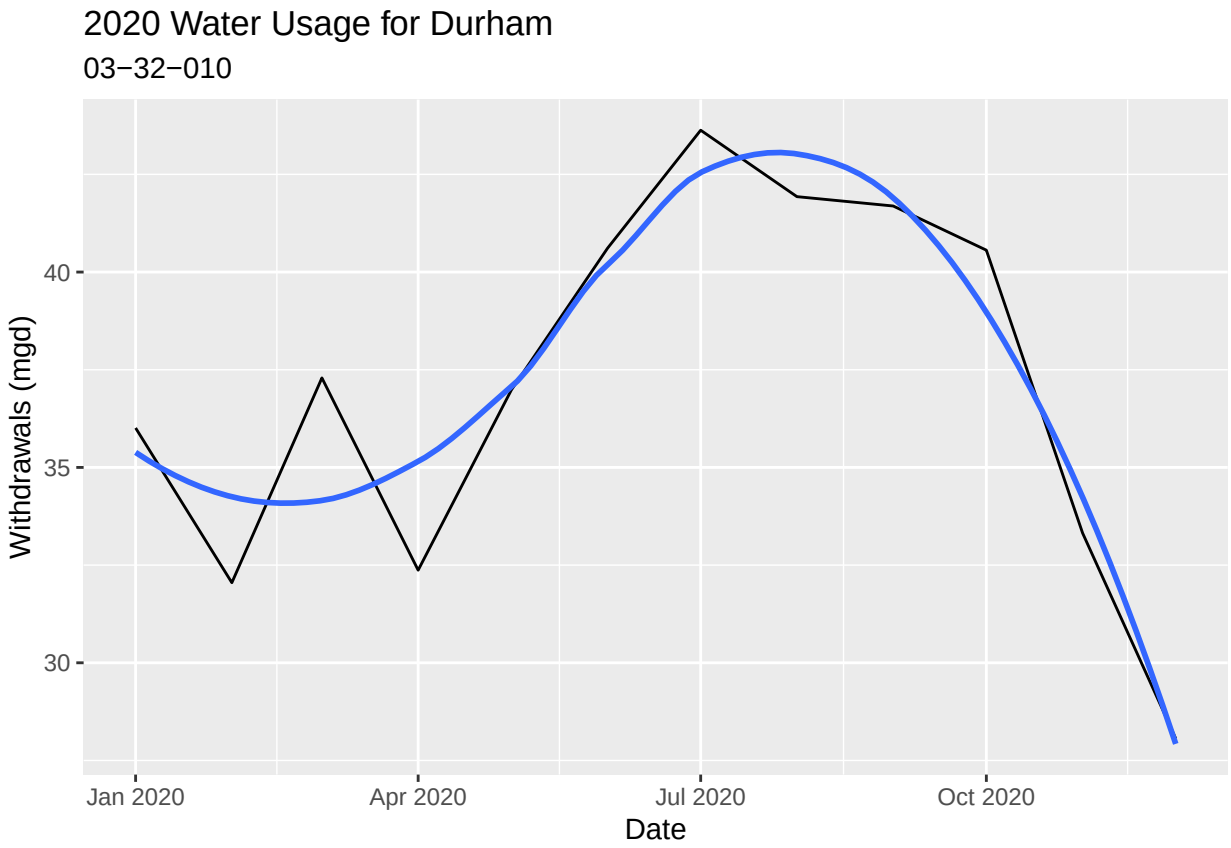
   TIP: Use `rep()` to repeat a value when creating a dataframe.

   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

#create month vector
month <- c("Jan", "May", "Sep", "Feb", "Jun",
           "Oct", "Mar", "Jul", "Nov", "Apr", "Aug", "Dec")

#create dataframe
Durham_withdrawals <- data.frame("Month" = month,
                                 "Year" = rep(2020,12),
                                 "Max_Withdrawals_mgd" = as.numeric(
                                   max.withdrawals.mgd))

#mutate dataframe
Durham_withdrawals <-Durham_withdrawals %>%
  mutate(Ownership = !!ownership,
         PWSID = !!pwsid,
         Water.System =!!water.system.name,
        Date = my(paste(Month,"-",Year)))


#5
Durham.withdrawals.plot <-
  ggplot(Durham_withdrawals, aes(x=Date, y=Max_Withdrawals_mgd))+
           geom_line()+
           geom_smooth(method="loess",se=FALSE) +
           labs(title = paste("2020 Water Usage for", water.system.name),
         subtitle = pwsid,
         y = "Withdrawals (mgd)",
         x="Date")

Durham.withdrawals.plot
```

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



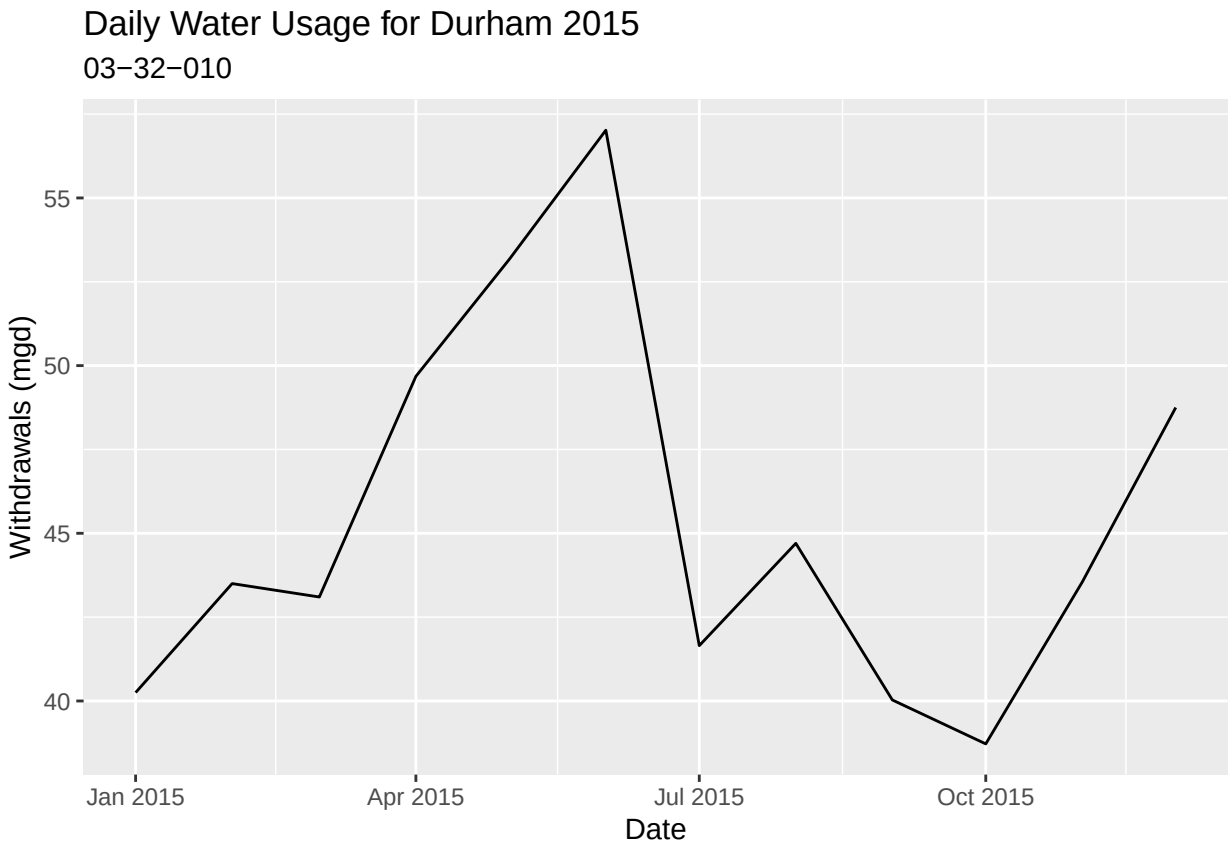
6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site scraped**.

```
#6.
#Construct the scraping web address, i.e. its URL
the_base_url <- 'https://www.ncwater.org/WUDC/app/LWSP/report.php?'
the_PWSID <- '03-32-010'
the_year <- 2020

PWSID.year.scrape.it <-function(the_year, the_PWSID){
  webpage <- read_html(paste0(
  "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=",
  the_PWSID, "&year=", the_year
  ))


#Set tags

water.system.tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'
PWSID.tag <- 'td tr:nth-child(1) td:nth-child(5)'
ownership.tag <-'div+ table tr:nth-child(2) td:nth-child(4)'
max.withdrawals.mgd.tag <- 'th~ td+ td'
month <- c("Jan", "May", "Sep", "Feb", "Jun", "Oct", "Mar", "Jul",
           "Nov", "Apr", "Aug", "Dec")

#Scrape the data items
the.water.system <- webpage %>% html_nodes(water.system.tag) %>% html_text()
```

```
the.PWSID.scrape<- webpage %>%   html_nodes(PWSID.tag) %>%  html_text()
the.ownership <- webpage %>% html_nodes(ownership.tag) %>% html_text()
the_withdrawals <- webpage %>% html_nodes(max.withdrawals.mgd.tag) %>% html_text()

#dataframe

withdrawals.dataframe <- data.frame("Month" = month,
      "Year" = rep(the_year,12),
      "Max_Withdrawals_mgd" = as.numeric(the_withdrawals)) %>%
  mutate(Ownership = !!the.ownership,
         PWSID = !!the.PWSID.scrape,
         Water.System =!!the.water.system,
        Date = as.Date(my(paste(Month,"-",Year))))
}
```

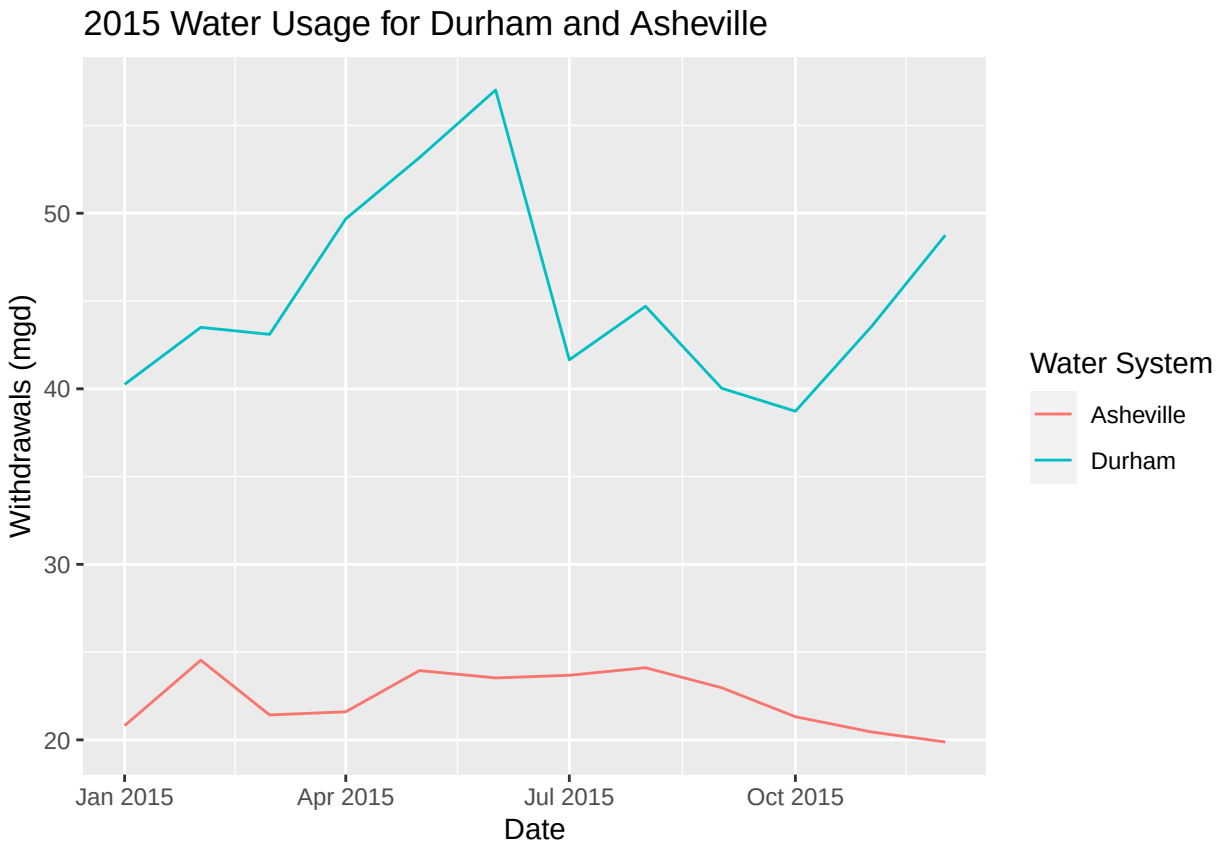
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID=‘03-32-010’) for each month in 2015

```
#7

Durham.2015 <- PWSID.year.scrape.it(2015,'03-32-010')

Durham.2015.plot <-
  ggplot(Durham.2015)+
  geom_line(aes(x = Date, y = Max_Withdrawals_mgd))+
  labs(x = "Date", y = "Withdrawals (mgd)",
       title = "Daily Water Usage for Durham 2015",
       subtitle = pwsid)

Durham.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 the Asheville to Durham's water withdrawals.

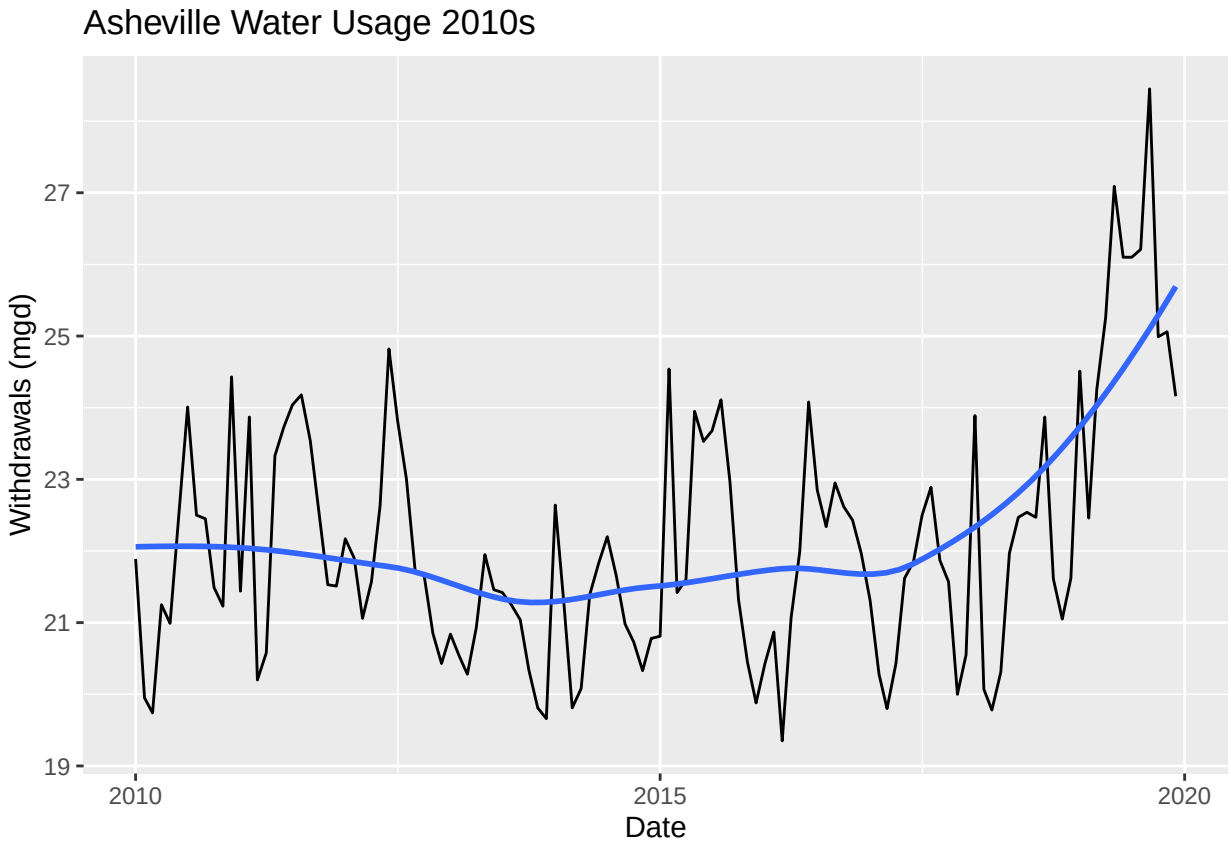
```
#8

#Asheville data
Asheville.2015 <- PWSID.year.scrape.it(2015, "01-11-010")

#combine data
data.combined <- rbind(Durham.2015, Asheville.2015)

#Asheville Durham plot
Asheville.Durham.plot <- ggplot(
  data.combined)+
  geom_line(aes(x = Date, y = Max_Withdrawals_mgd, color = Water.System)) +
  labs(x = "Date", y = "Withdrawals (mgd)", title =
"2015 Water Usage for Durham and Asheville",
       color = "Water System")+
 theme(legend.position="right")


Asheville.Durham.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.

```
#9

#defining object values
the_year = rep(2010:2019)
the_PWSID = '01-11-010'

#Asheville function
Asheville.2010s <-lapply(X = the_year,
                  FUN = PWSID.year.scrape.it,
                  the_PWSID=the_PWSID)

#Asheville dataframe
Asheville.2010s.df <- bind_rows(Asheville.2010s)

#Asheville plot
Asheville.2010s.plot <-
  ggplot(Asheville.2010s.df, aes(x = Date, y = Max_Withdrawals_mgd))+
           geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title =
         paste("Asheville Water Usage 2010s"),
       y="Withdrawals (mgd)",
       x="Date")

Asheville.2010s.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? From the plot, it appears that Asheville does have an increasing trend in water usage over time. Additionally, it looks like there is a seasonal component to this data, and that water usage varies throughout each year depending on the month/season. We would need to conduct a time-series analysis on this data to further determine if there is a seasonal component.