Assignment 09: Data Scraping

Tasha Griffiths

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. check directory
getwd()
```

```
## [1] "C:/Users/Tasha Griffiths/Documents/Duke Year 1/Spring 22 Classes/Environmental Data Analytics/G
```

```
#2. load libraries
library(tidyverse)
library(lubridate)
library(viridis)
library(rvest)
library(dataRetrieval)
```

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

```
library(tidycensus)
```

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

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

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
webpageNC_DEQ <-
  read_html('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2020')
webpageNC_DEQ
```

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

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

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

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

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

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

```
max.withdrawals.mgd <- webpageNC_DEQ %>%
  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 convert max withdrawls to dataframe

#need to pull month information from web table
max.withdrawls.month <- webpageNC_DEQ %>%
  html_nodes(".fancy-table:nth-child(31) tr+ tr th") %>%
  html_text()
max.withdrawls.month
```

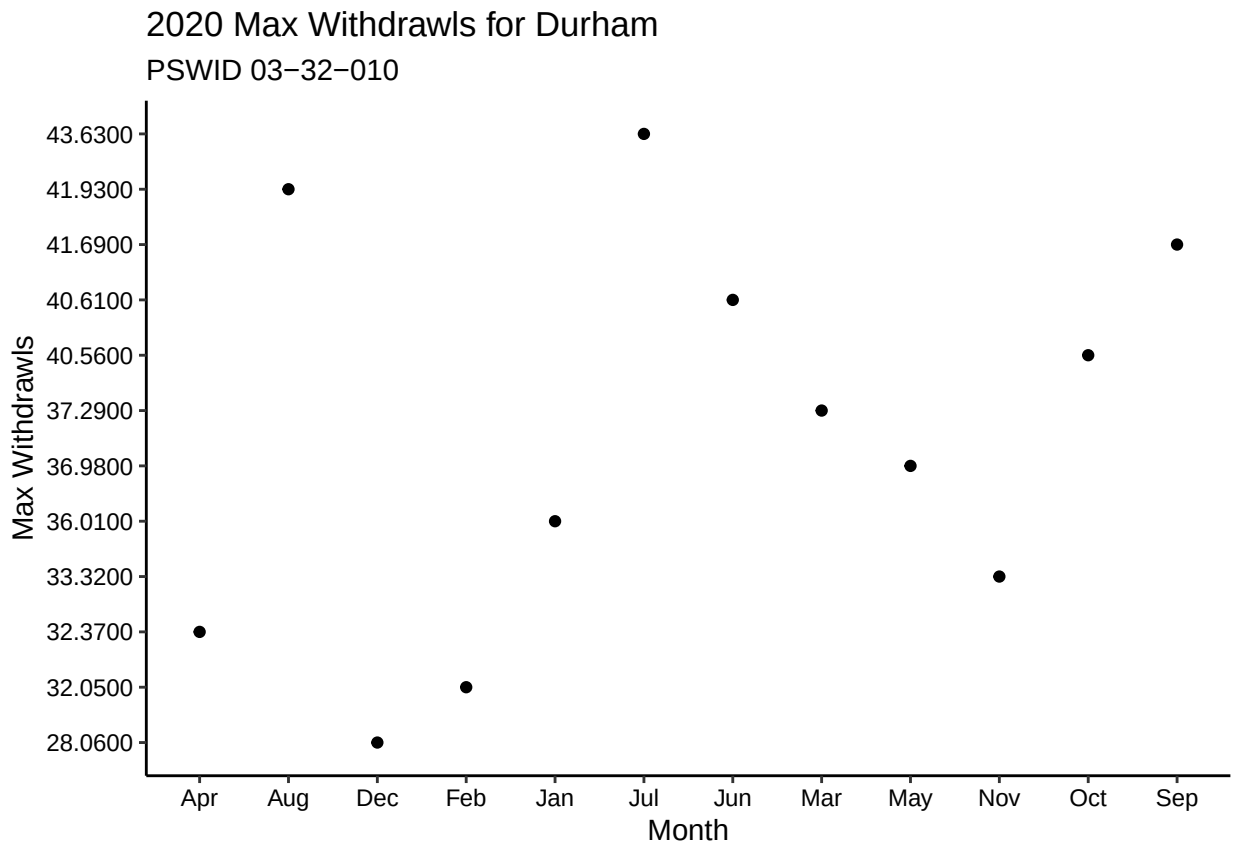
```
##  [1] "Jan" "May" "Sep" "Feb" "Jun" "Oct" "Mar" "Jul" "Nov" "Apr" "Aug" "Dec"
```

```
#create a year and date variable to add to dataframe
max.withdrawls.year <- 2020
Date <- as.Date(my(paste(max.withdrawls.month, "-", max.withdrawls.year)))
class (Date)
```

```
## [1] "Date"
```

```
#create the dataframe
the_df <- data.frame(
"Water System Name" = water.system.name,
"PSWID" = pswid,
"Ownership" = ownership,
"Max Withdrawls Total" = as.numeric(max.withdrawals.mgd),
"Max Withdrawls Month" = max.withdrawls.month,
"Max Withdrawls Year" = max.withdrawls.year,
"Date" = as.Date(my(paste(max.withdrawls.month, "-", max.withdrawls.year)))
)


#5 Plot couldn't get a line to work so used points
ggplot(the_df,aes(x=max.withdrawls.month,y=max.withdrawals.mgd)) +
  geom_point() +
  labs(title = paste("2020 Max Withdrawls for", water.system.name),
       subtitle = paste("PSWID", pswid),
       y="Max Withdrawls",
       x="Month")
```



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 a function to scrape data for any PSWID and Year

Withdrawls.scrape.function <- function(any_year, pswid_number){
  #fetch website
  the_url <- read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pswid='
                    ,pswid_number, '&year=',any_year))
  print(the_url)

  #scrape data
  water.system.name.new <- "div+ table tr:nth-child(1) td:nth-child(2)"
  pswid.new <- "td tr:nth-child(1) td:nth-child(5)"
  ownership.new <- "div+ table tr:nth-child(2) td:nth-child(4)"
  max.withdrawals.mgd.new <- "th~ td+ td"
  max.withdrawls.month.new <- ".fancy-table:nth-child(31) tr+ tr th"
  max.withdrawls.year.new <- any_year


  web.system.name <- the_url %>% html_nodes(water.system.name.new) %>% html_text()
  web.pswid <- the_url %>% html_nodes(pswid.new) %>% html_text()
  web.ownership <- the_url %>% html_nodes(ownership.new) %>% html_text()
  web.max.withdrawls <- the_url %>% html_nodes(max.withdrawals.mgd.new) %>% html_text()
  web.withdrawls.month <- the_url %>% html_nodes(max.withdrawls.month.new) %>% html_text()

  #convert to dataframe
  new.dataframe <- data.frame("Water System Name" = web.system.name,
                              "PSWID" = web.pswid,
                              "Ownership" = web.ownership,
                              "Max_Withdrawls_Total" = as.numeric(web.max.withdrawls),
                              "Max_Withdrawls_Month" = web.withdrawls.month,
                              "Max_Withdrawls_Year" = max.withdrawls.year.new #,
                              #"Date" = as.Date(my(paste(web.max.withdrawls, "-",
                                                        #max.withdrawls.year.new)))
)

  #wrangle new dataframe
  new.dataframe <- new.dataframe %>% mutate
  ("Date_New" = my(paste(Max_Withdrawls_Month, "-", Max_Withdrawls_Year)))

  #show the dataframe
  return(new.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 scape data for new year
#durham.2015.withdrawls <- Withdrawls.scrape.function('2015','03-32-010')
#view(durham.2015.withdrawls)

#plot 1
#ggplot(durham.2015.withdrawls) +
  #geom_point(aes(x=Date_New,y=Max_Withdrawls_Total)) +
  #labs(title = paste("2015 Max Withdrawls for",
```

```
                    #durham.2015.withdrawls$water.system.name),
       #subtitle = paste("PSWID", durham.2015.withdrawls$PSWID),
       #y="Max Withdrawls",
       #x="Month")
```

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 scape data for new location
#asheville.2015.withdrawls <- Withdrawls.scrape.function('2015','01-11-010')
#view(asheville.2015.withdrawls)

#won't pull for asheville. Getting a error
#Error in data.frame(`Water System Name` = web.system.name, PSWID = web.pswid, :
#arguments imply differing number of rows: 1, 12, 0

#plot
#plot <- ggplot(asheville.2015.withdrawls) +
  #geom_point(aes(x=Date,y=Max_Withdrawls_Total)) +
  #abs(title = paste("2015 Max Withdrawls for", water.system.name),
       #subtitle = paste("PSWID", durham.2015.withdrawls$PSWID),
       #y="Max Withdrawls",
       #x="Month")

#plot2 <- ggplot() +
#geom_line(data=durham.2015.withdrawls,
#aes(x=Date_New, y=Max_Withdrawls_Total), color='green') +
#geom_line(data=asheville.2015.withdrawls,
#aes(x=Date_New, y=Max_Withdrawls_Total), color='yelow') +
  #labs(title = paste("2015 Max Withdrawls for",
#durham.2015.withdrawls$water.system.name,
#asheville.2015.withdrawls$water.system.name),
       #subtitle = paste("PSWID", durham.2015.withdrawls$PSWID,
#asheville.2015.withdrawls$PSWID),
       #y="Max Withdrawls",
       #x="Month")


#after spending over 12 hours on this assignment, re-reviewing lessons, lab notes, and slack I was unab
```

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 scrape 2010 to 2019
#any_year = seq(2010,2020)
#pswid = '01-11-010'

#plot
#plot3 <- ggplot(durham.2015.withdrawls) +
  #geom_point(aes(x=max.withdrawls.month.new,y=max.withdrawals.mgd.new)) +
  #labs(title = paste("2020 Max Withdrawls for", water.system.name.new),
```

```
  #subtitle = paste("PSWID", pswid.new),
  #y="Max Withdrawls",
  #x="Month")
```

Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time?