Assignment 09: Data Scraping

Sophia Bryson

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

```
getwd()
```

```
## [1] "Z:/ENV872/Environmental_Data_Analytics_2022/Assignments"
```

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

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

```
library(wesanderson)

theme_set(theme_minimal())
```

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 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.)

```
# NOTES FROM JOHN: Scrape 2020 and max withdrawls
#2
url <- "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2020"
webpage <- read_html(url)
```

3. The data we want to collect are listed below:

- From the "1. System Information" section:
- Water system name
- PWSID
- 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 <- webpage %>% html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>% html_text
pwsid <- webpage %>% html_nodes('td tr:nth-child(1) td:nth-child(5)') %>% html_text()
ownership <- webpage %>% html_nodes('div+ table tr:nth-child(2) td:nth-child(4)') %>% html_text()
max.withdrawals.mgd <- webpage %>% html_nodes('th~ td+ td') %>% html_text()
max.withdrawl.months <- webpage %>% html_nodes('.fancy-table:nth-child(31) tr+ tr th') %>% html_text()
```

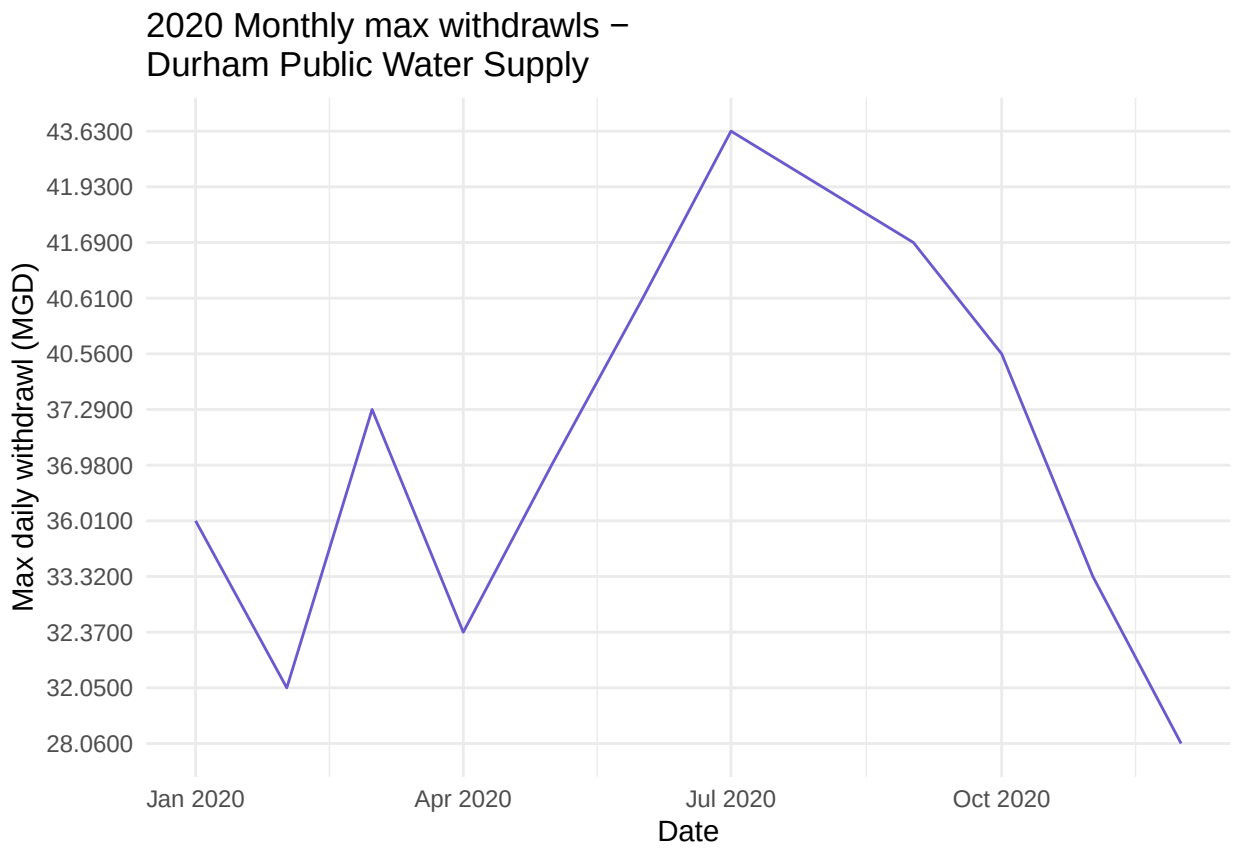
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   > 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
DurhamWithdrawls <- data.frame(SystemName = c(rep(water.system.name, length(max.withdrawals.mgd))),
                               PWSID = c(rep(pwsid, length(max.withdrawals.mgd))),
                               Ownership = c(rep(ownership, length(max.withdrawals.mgd))),
                               MaxWithdrawls_mgd = max.withdrawals.mgd,
                               Month = max.withdrawl.months) %>%
                      mutate(Date = as.Date(paste0(Month, "-1-2020"), format = "%b-%d-%Y")) %>%
                      select(- Month) %>%
                      arrange(Date)

#5
ggplot(DurhamWithdrawls, aes(x = Date, y = MaxWithdrawls_mgd, group = SystemName)) +
  geom_line(color = "slateblue") +
  labs(y = "Max daily withdrawl (MGD)", title = "2020 Monthly max withdrawls - \nDurham Public Water Su
```



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.

#Manual month order because it's not working for Asheville
monthOrder <- max.withdrawl.months

scrapeNCDEQ <- function(PWSID, year) {
  url <- paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=', PWSID, '&year=', as.characte
  webpage <- read_html(url)

  water.system.name <- webpage %>% html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>% html_te
  pwsid <- webpage %>% html_nodes('td tr:nth-child(1) td:nth-child(5)') %>% html_text()
  ownership <- webpage %>% html_nodes('div+ table tr:nth-child(2) td:nth-child(4)') %>% html_text()
  max.withdrawals.mgd <- webpage %>% html_nodes('th~ td+ td') %>% html_text()
  max.withdrawl.months <- monthOrder

  Withdrawls <- data.frame(SystemName = c(rep(water.system.name, length(max.withdrawals.mgd))),
                           PWSID = c(rep(pwsid, length(max.withdrawals.mgd))),
                           Ownership = c(rep(ownership, length(max.withdrawals.mgd))),
                           MaxWithdrawls_mgd = max.withdrawals.mgd,
                           Month = max.withdrawl.months) %>%
                mutate(Date = as.Date(paste0(Month, "-1-", year), format = "%b-%d-%Y")) %>%
                select(- Month) %>%
                arrange(Date)
```

```
  return(Withdrawls)
}
```

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

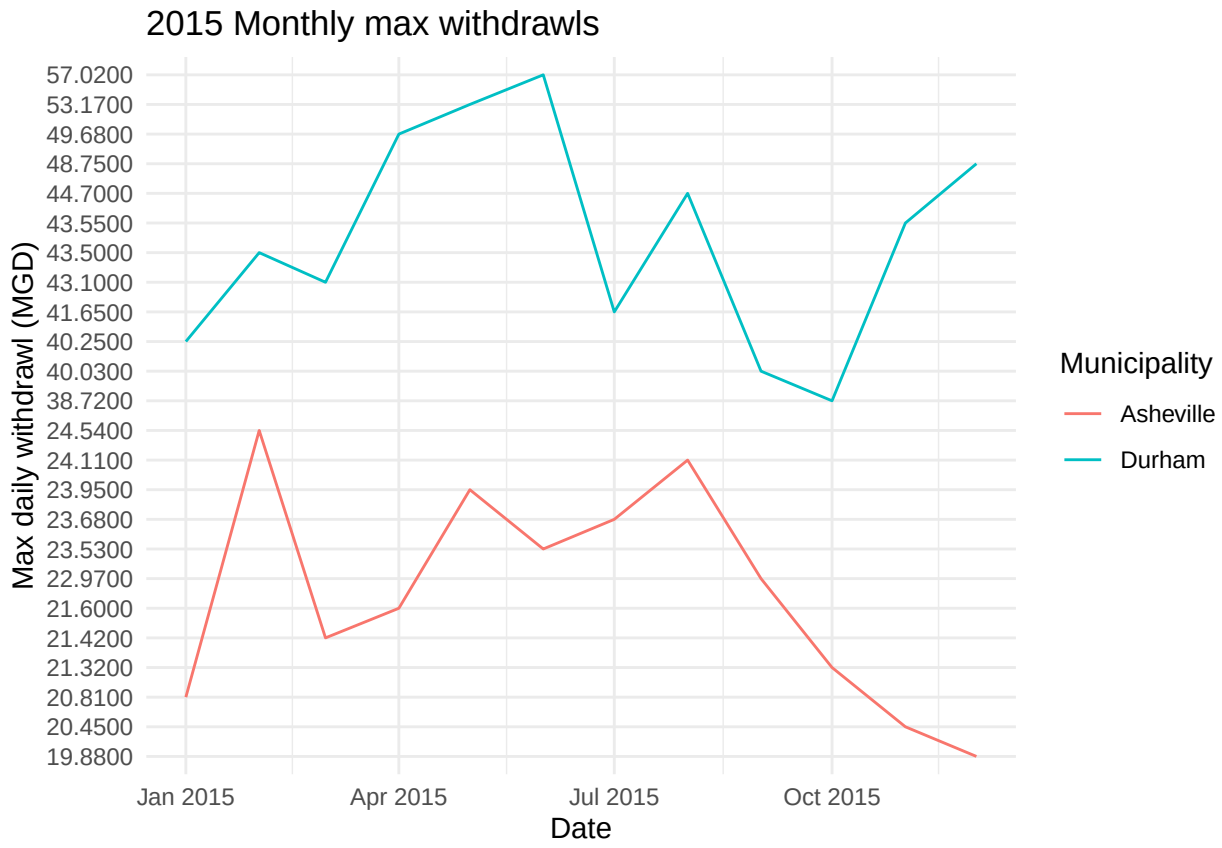
```
#7
Durham2015 <- scrapeNCDEQ('03-32-010', 2015)
```

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 <- scrapeNCDEQ('01-11-010', 2015)

Withdrawls2015 <- rbind(Durham2015, Asheville2015)

ggplot(Withdrawls2015, aes(x = Date, y = MaxWithdrawls_mgd, group = SystemName, color = SystemName)) +
  geom_line() +
  labs(y = "Max daily withdrawl (MGD)", title = "2015 Monthly max withdrawls",
       color = "Municipality")
```



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.

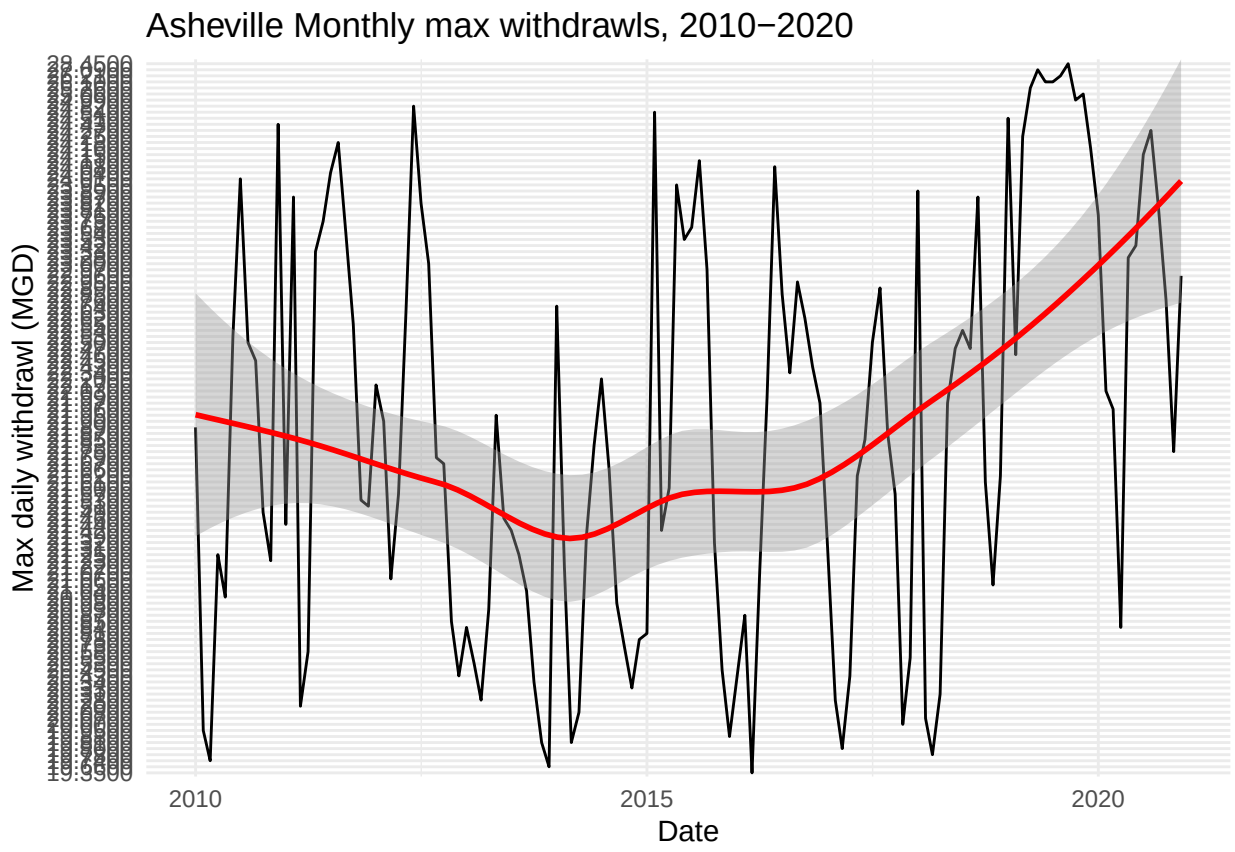
```
#9
years <- seq(2010, 2020, 1)
```

```
year.dfs <- lapply(X = years,
                   FUN = scrapeNCDEQ,
                   PWSID = '01-11-010')

years.df <- bind_rows(year.dfs)

ggplot(years.df, aes(x = Date, y = MaxWithdrawls_mgd, group = SystemName)) +
  geom_line() +
  geom_smooth(color = "red") +
  labs(y = "Max daily withdrawl (MGD)", title = "Asheville Monthly max withdrawls, 2010-2020")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
#SORRY THE Y-AXIS IS HIDEOUS. I PROMISE I TRIED TO FIX IT BUT IT KEPT THROWING ERRORS.
```

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

Visually examining the plot suggests taht Asheville's water usage was decreasing from 2010 through 2014 and has subsequently increasing fairly steadily up through 2020.