Assignment 09: Data Scraping

Amanda Booth

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 working directory

getwd()
```

```
## [1] "/Users/amandabooth/Documents/GitHub/Environmental_Data_Analytics_2022/Assignments"
```

```
#Load packages

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

#Build theme
mytheme <- theme_linedraw(base_size = 10) +
  theme(axis.text = element_text(color = "rosybrown"),
```

```
        legend.position = "bottom")

#Set as default theme
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 2021 to 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')
```

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.)

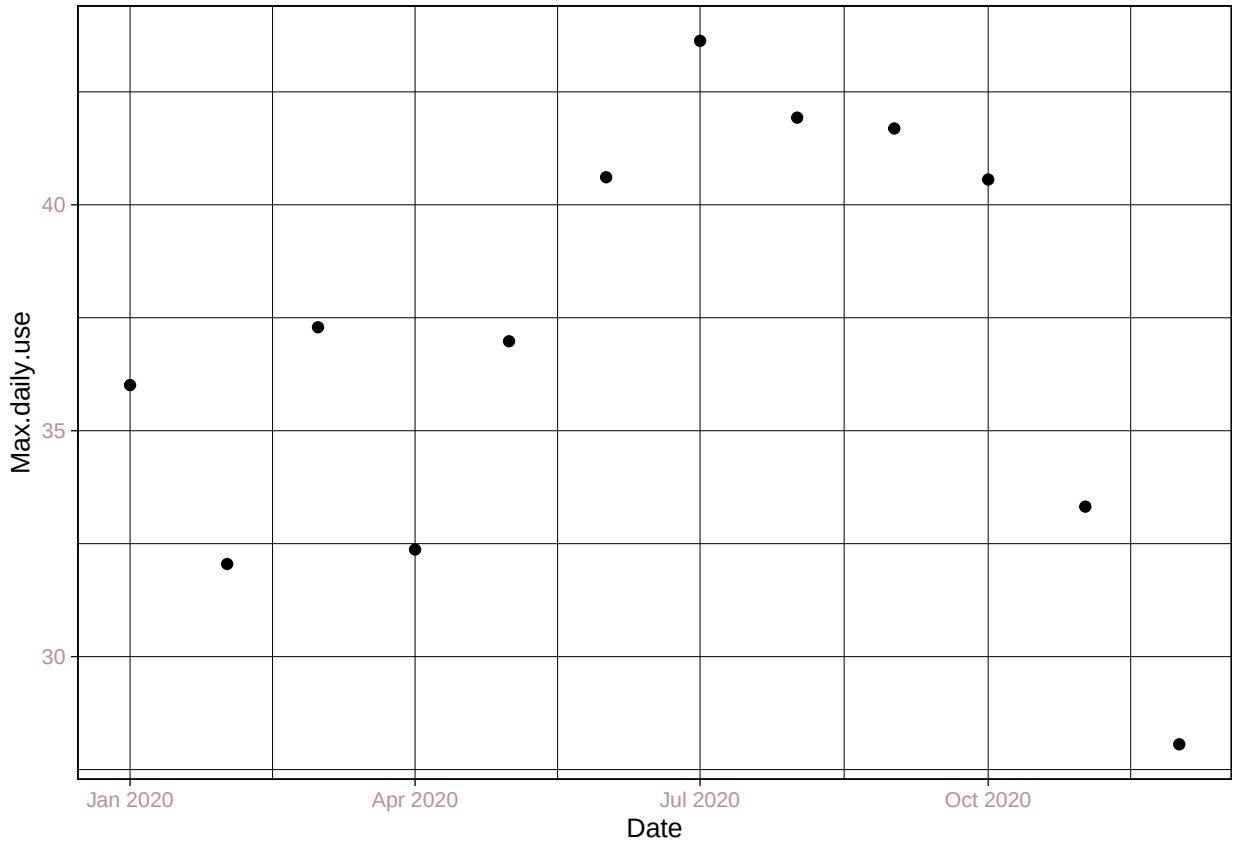
   TIP: Use `rep()` to repeat a value when creating a dataframe.

   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

Durham_df <- data.frame(
  "Water system name" = water.system.name,
  "Ownership" = ownership,
  "PSWID" = pswid,
  "Max daily use" = as.numeric(max.withdrawals.mgd),
  "Month" = c(01, 05, 09, 02, 06, 10, 03, 07, 11, 4, 08, 12),
  "Year" = 2020)

Durham_df <- arrange(Durham_df, Month)

Durham_df <- Durham_df %>%
  mutate(Date = my(paste(Month,"-",Year)))

Durham_df$Month <- month.abb[Durham_df$Month]

#5
```

```r
Plot1 <- ggplot(Durham_df) +
  geom_point(aes(x = Date, y = Max.daily.use))

print(Plot1)
```



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**.

```r
#6.

#Create function
scrape.it <- function(Year, pswid){

#Retrieve content
website <- read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',
                                  pswid, '&year=', Year))


#Set tags

water.system.tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'
pswid.tag <- 'td tr:nth-child(1) td:nth-child(5)'
ownership.tag <-  'div+ table tr:nth-child(2) td:nth-child(4)'
max.use.tag <- 'th~ td+ td'
```

```
#Scrape data
the_water.system <- website %>% html_nodes(water.system.tag) %>% html_text()
the_pswid <- website %>%   html_nodes(pswid.tag) %>%  html_text()
the_ownership <- website %>% html_nodes(ownership.tag) %>% html_text()
the_max.use <- website %>% html_nodes(max.use.tag) %>% html_text()

#Create dataframe
df_withdrawals <- data.frame("Month" = c(01, 05, 09, 02, 06, 10, 03, 07, 11, 4, 08, 12),
                             "Year" = rep(Year,12),
                             "Max.daily.use" = as.numeric(the_max.use)) %>%
    mutate(Water_system = !!the_water.system,
           PSWID = !!the_pswid,
           Ownership = !!the_ownership,
           Date = my(paste(Month,"-",Year)))


#Return dataframe
return(df_withdrawals)

}
```

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_df <- scrape.it(2015,'03-32-010')
view(Durham2015_df)
```

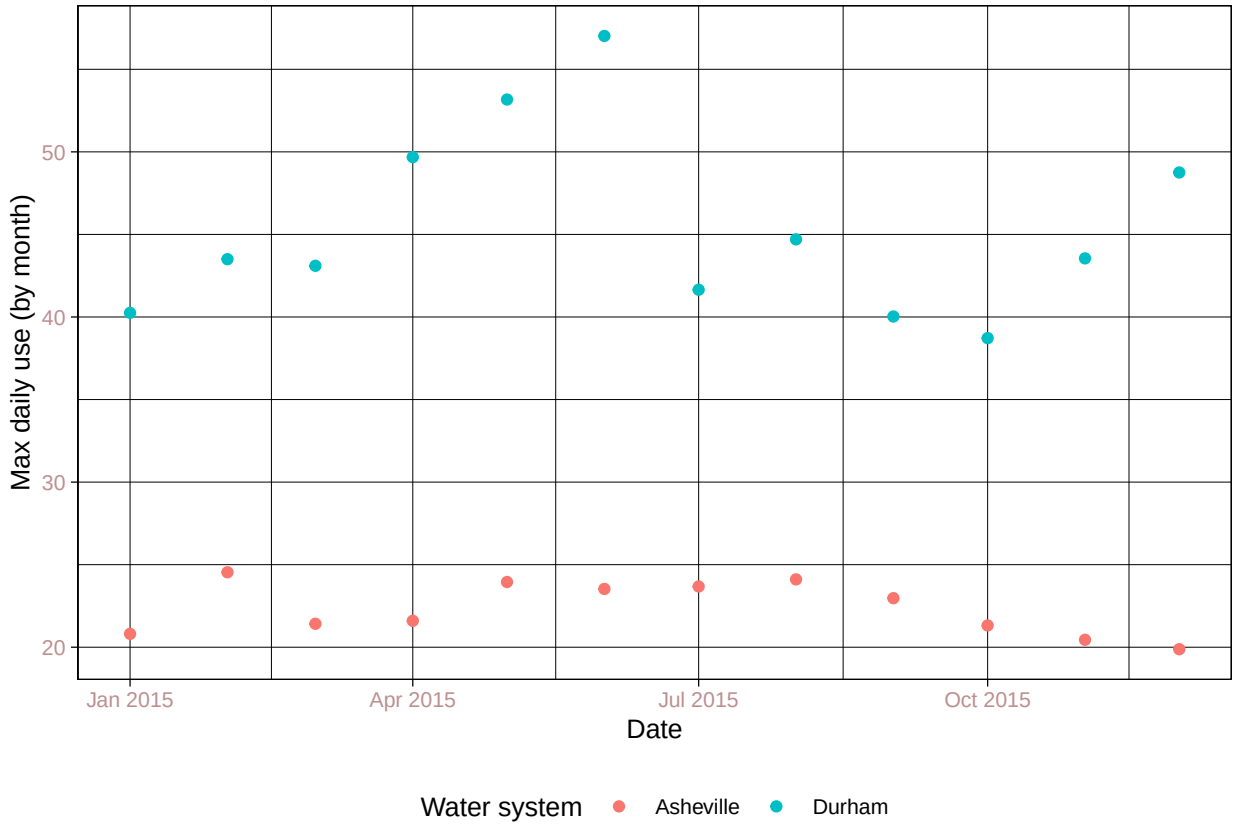
8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares the Asheville to Durham's water withdrawals.

```
#8

#Scrape Asheville
Asheville2015_df <- scrape.it(2015,'01-11-010')
view(Asheville2015_df)

#Combine data

df_2015 <-  bind_rows(Durham2015_df, Asheville2015_df)

  merge(x = Durham2015_df, y = Asheville2015_df, by = "Date")
```

```
##          Date Month.x Year.x Max.daily.use.x Water_system.x   PSWID.x
## 1  2015-01-01       1   2015           40.25         Durham 03-32-010
## 2  2015-02-01       2   2015           43.50         Durham 03-32-010
## 3  2015-03-01       3   2015           43.10         Durham 03-32-010
## 4  2015-04-01       4   2015           49.68         Durham 03-32-010
## 5  2015-05-01       5   2015           53.17         Durham 03-32-010
```

```
## 6  2015-06-01       6   2015           57.02         Durham 03-32-010
## 7  2015-07-01       7   2015           41.65         Durham 03-32-010
## 8  2015-08-01       8   2015           44.70         Durham 03-32-010
## 9  2015-09-01       9   2015           40.03         Durham 03-32-010
## 10 2015-10-01      10   2015           38.72         Durham 03-32-010
## 11 2015-11-01      11   2015           43.55         Durham 03-32-010
## 12 2015-12-01      12   2015           48.75         Durham 03-32-010
##     Ownership.x Month.y Year.y Max.daily.use.y Water_system.y   PSWID.y
## 1  Municipality       1   2015           20.81      Asheville 01-11-010
## 2  Municipality       2   2015           24.54      Asheville 01-11-010
## 3  Municipality       3   2015           21.42      Asheville 01-11-010
## 4  Municipality       4   2015           21.60      Asheville 01-11-010
## 5  Municipality       5   2015           23.95      Asheville 01-11-010
## 6  Municipality       6   2015           23.53      Asheville 01-11-010
## 7  Municipality       7   2015           23.68      Asheville 01-11-010
## 8  Municipality       8   2015           24.11      Asheville 01-11-010
## 9  Municipality       9   2015           22.97      Asheville 01-11-010
## 10 Municipality      10   2015           21.32      Asheville 01-11-010
## 11 Municipality      11   2015           20.45      Asheville 01-11-010
## 12 Municipality      12   2015           19.88      Asheville 01-11-010
##     Ownership.y
## 1  Municipality
## 2  Municipality
## 3  Municipality
## 4  Municipality
## 5  Municipality
## 6  Municipality
## 7  Municipality
## 8  Municipality
## 9  Municipality
## 10 Municipality
## 11 Municipality
## 12 Municipality
```

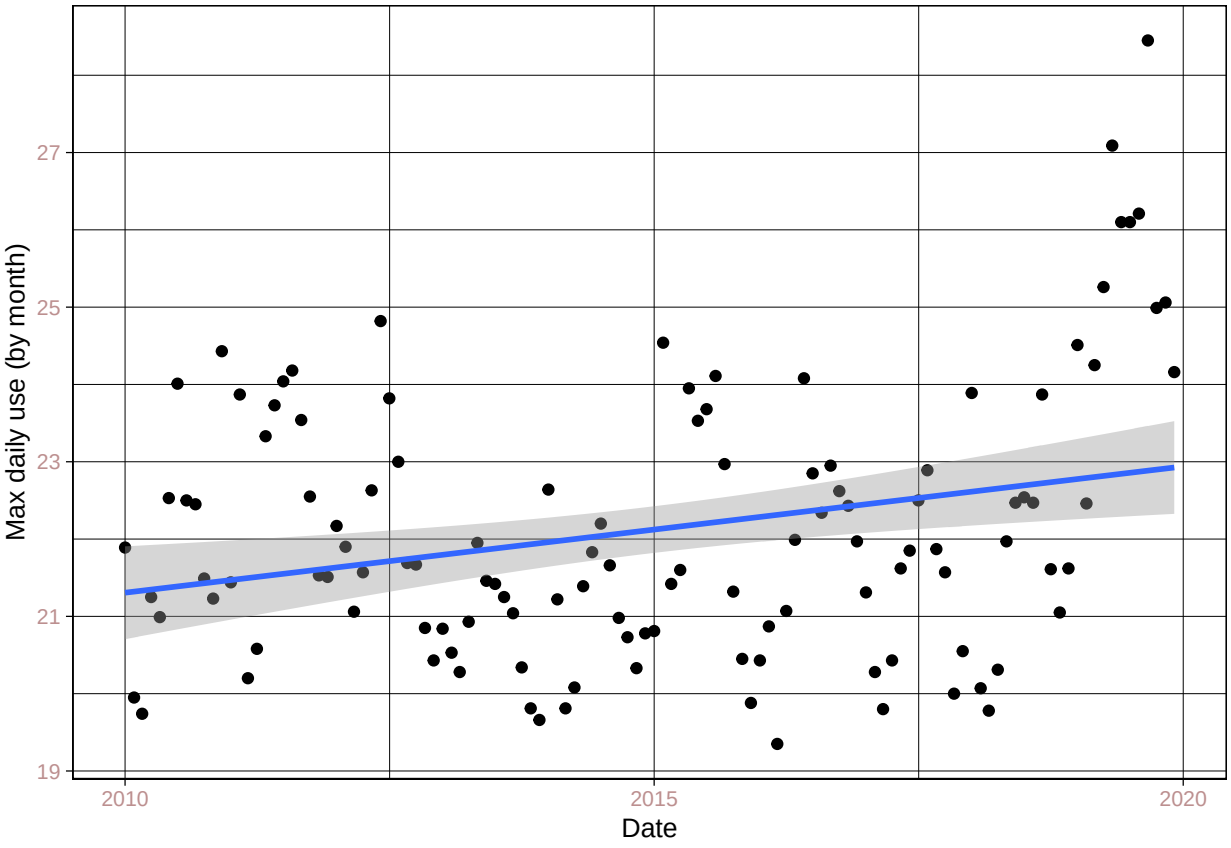
```r
#Plot data

Plot2 <- ggplot(df_2015) +
  geom_point(aes(x = Date, y = Max.daily.use, color = Water_system)) +
  labs(y = "Max daily use (by month)", color = "Water system")

print(Plot2)
```



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 2010-2019

the_years = rep(2010:2019)
asheville = '01-11-010'
asheville_dfs <- lapply(X = the_years,
                  FUN = scrape.it,
                  pswid = asheville)

#Combine rows
asheville_df <- bind_rows(asheville_dfs)

#Plot

Plot3 <- ggplot(asheville_df, aes(x = Date, y = Max.daily.use)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(y = "Max daily use (by month)")

print(Plot3)
```

```
## `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?

Answer: Yes, the plot shows a trend that suggests there is an increase in water usage over time.