Assignment 09: Data Scraping

Benjamin Culberson

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

```r
#1
library(tidyverse)
library(rvest)
library(lubridate)

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 2020 Municipal Local Water Supply Plan (LWSP):

- Navigate to https://www.ncwater.org/WUDC/app/LWSP/search.php
- Change the date from 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
water_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:
- 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 <- water_webpage %>%
  html_nodes(’td tr:nth-child(1) td:nth-child(2)’) %>% html_text
water.system.name <- water.system.name[1]
pswid <- water_webpage %>%
  html_nodes(’td tr:nth-child(1) td:nth-child(5)’) %>% html_text
ownership <- water_webpage %>%
  html_nodes(’tr:nth-child(2) td:nth-child(4)’) %>% html_text
ownership <- ownership[1]
max.withdrawals.mgd <- water_webpage %>%
  html_nodes(’th~ td:nth-child(6) , th~ td:nth-child(3)’) %>% html_text
```

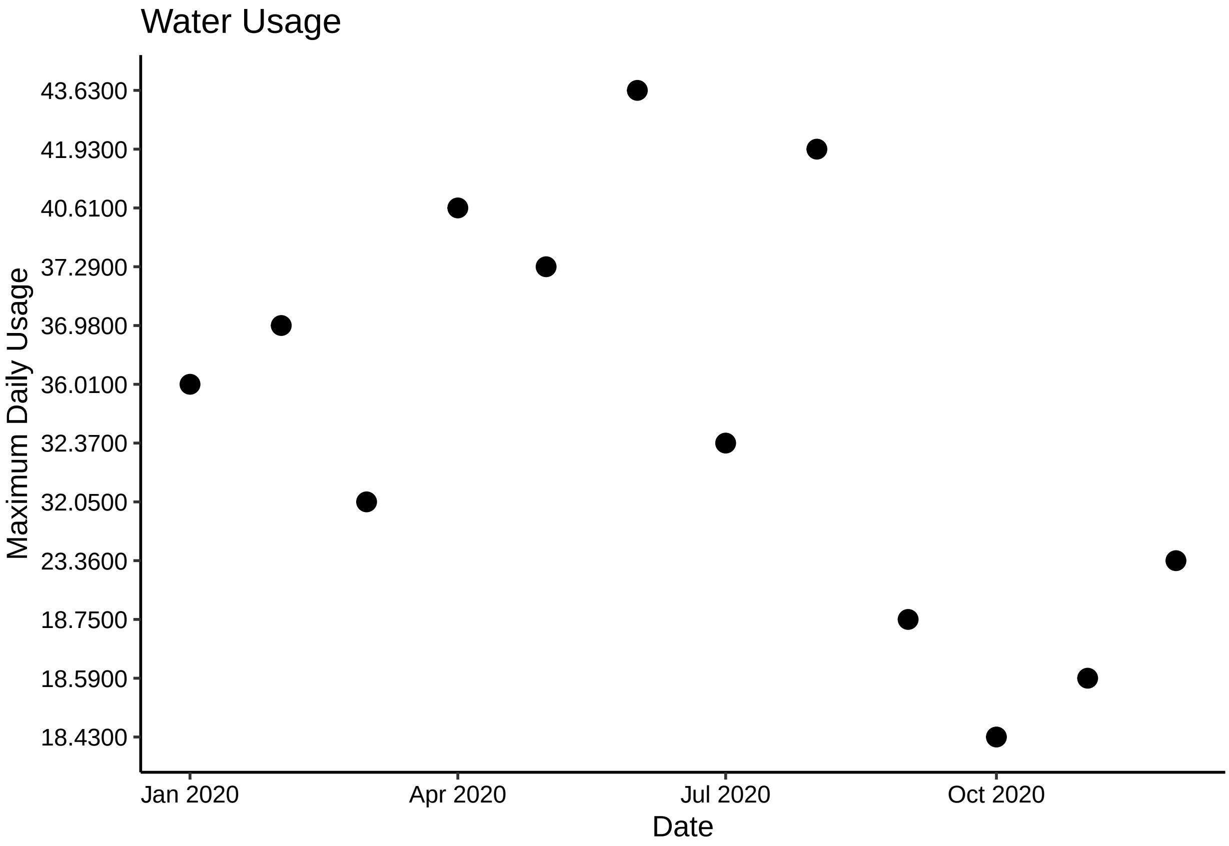
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   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

```r
#4
new_df <- data.frame(
  "Water System Name" = rep(water.system.name),
  "PSWID" = rep(pswid),
  "Ownership" = rep(ownership),
  "Maximum_Daily_Use" = max.withdrawals.mgd,
  month = rep(1:12),
  year = 2020
)

new_df <- new_df %>%
  mutate(Date = my(paste(month,"-",year)))

#5
ggplot(new_df, aes(x=Date, y=Maximum_Daily_Use)) +
  geom_point(size=3.0)+
  labs(title = "Water Usage", y = "Maximum Daily Usage")
```

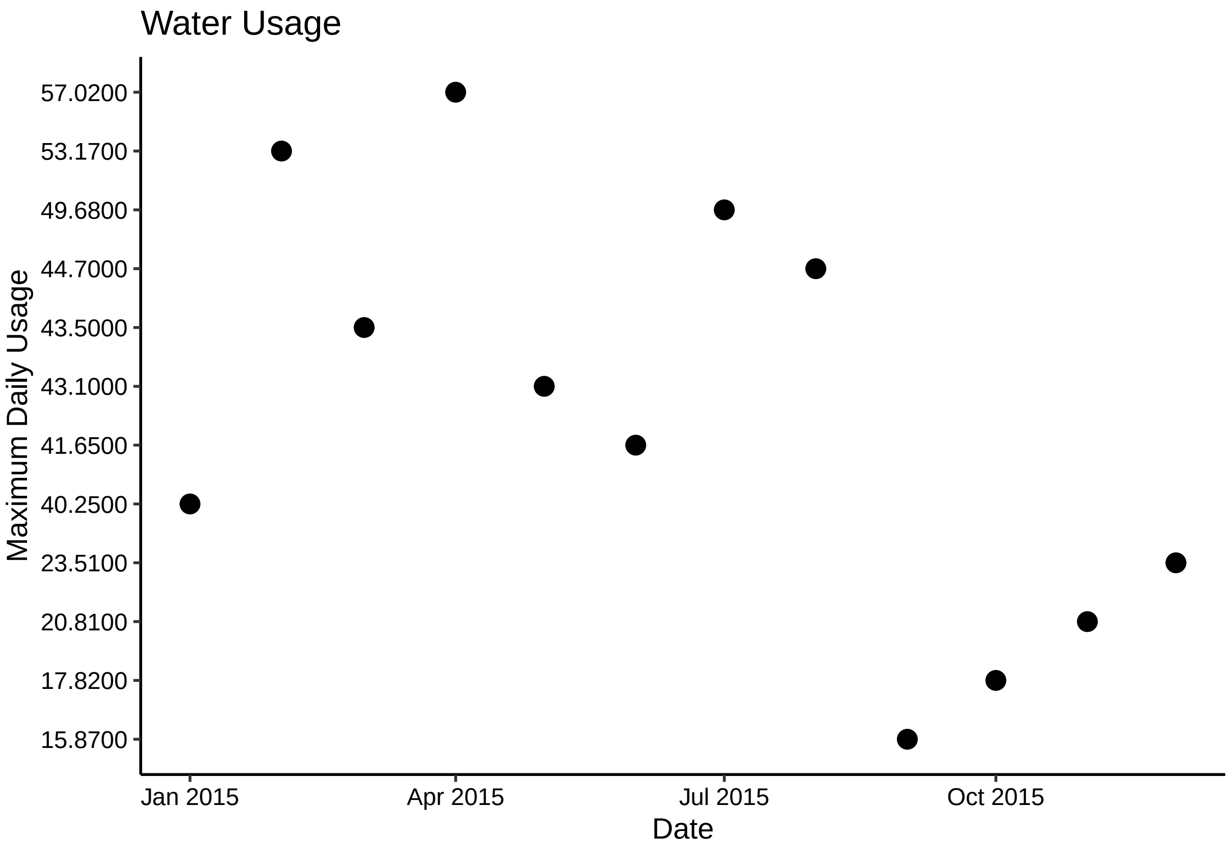


6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site scraped**.

```r
#6.
scrape.it <- function(the_year){
  #Get the proper url
  the_url <- ifelse(the_year==2021,
                    ’https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010’,
                    paste(’https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=’,
                          the_year,sep=’’))

  #1 Link to the web site using read_html
  water_website <- read_html(the_url)
  #print(water_website)

  #2&3 Locate elements and read their text attributes into variables
  water.system.name <- water_website %>%
    html_nodes(’td tr:nth-child(1) td:nth-child(2)’) %>% html_text
  water.system.name <- water.system.name[1]
  pswid <- water_website %>%
    html_nodes(’td tr:nth-child(1) td:nth-child(5)’) %>% html_text
  ownership <- water_website %>%
    html_nodes(’tr:nth-child(2) td:nth-child(4)’) %>% html_text
  ownership <- ownership[1]
  max.withdrawals.mgd <- water_website %>%
    html_nodes(’th~ td:nth-child(6) , th~ td:nth-child(3)’) %>% html_text

  #3 Construct a dataframe from the values
  new_df <- data.frame(
  "Water System Name" = rep(water.system.name),
  "PSWID" = rep(pswid),
  "Ownership" = rep(ownership),
  "Maximum_Daily_Use" = max.withdrawals.mgd,
  month = rep(1:12),
  year = the_year

  )
  new_df <- new_df %>%
    mutate(Date = my(paste(month,"-",year)))
  return(new_df)
}
```

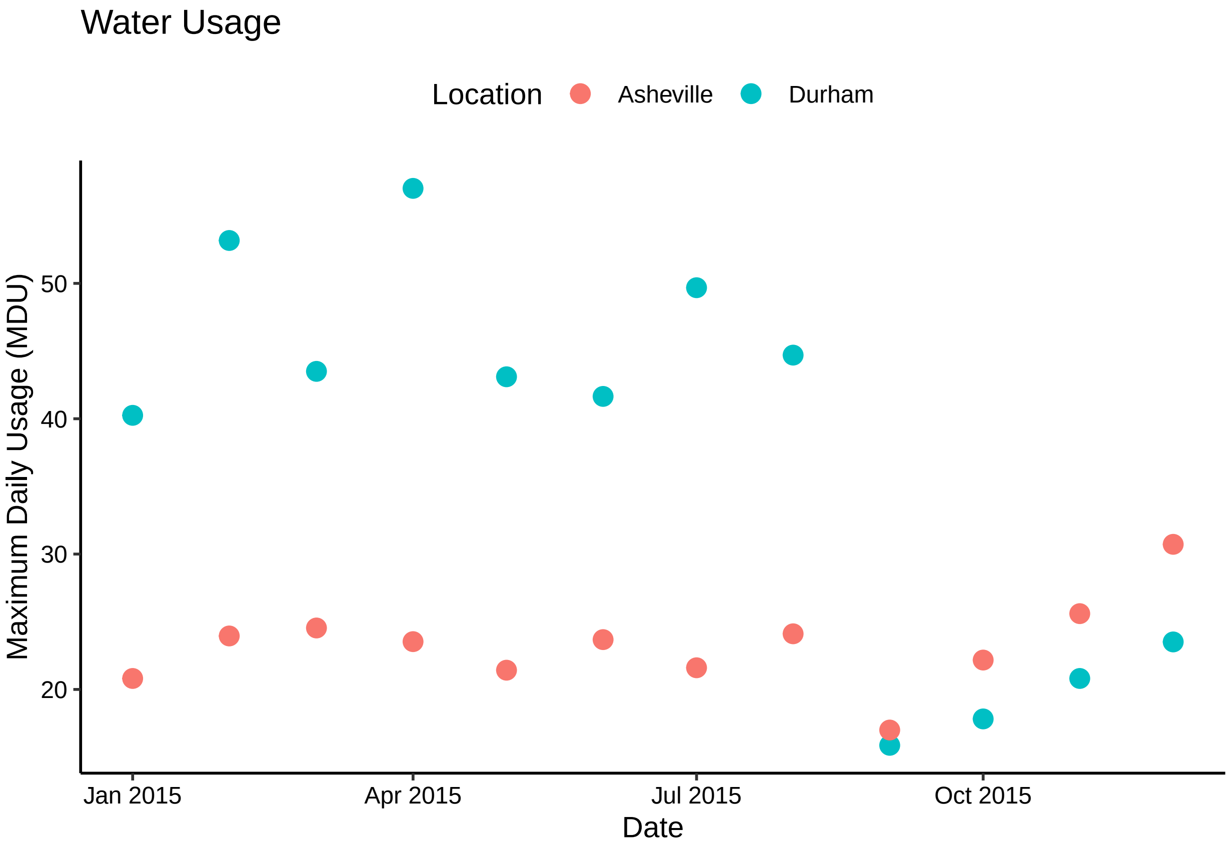
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID=‘03-32-010’) for each month in 2015

```
#7
Durham_2015_df <- scrape.it(2015)
ggplot(Durham_2015_df, aes(x=Date, y=Maximum_Daily_Use)) +
  geom_point(size=3.0)+
  labs(title = "Water Usage", y = "Maximum Daily Usage")
```



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.

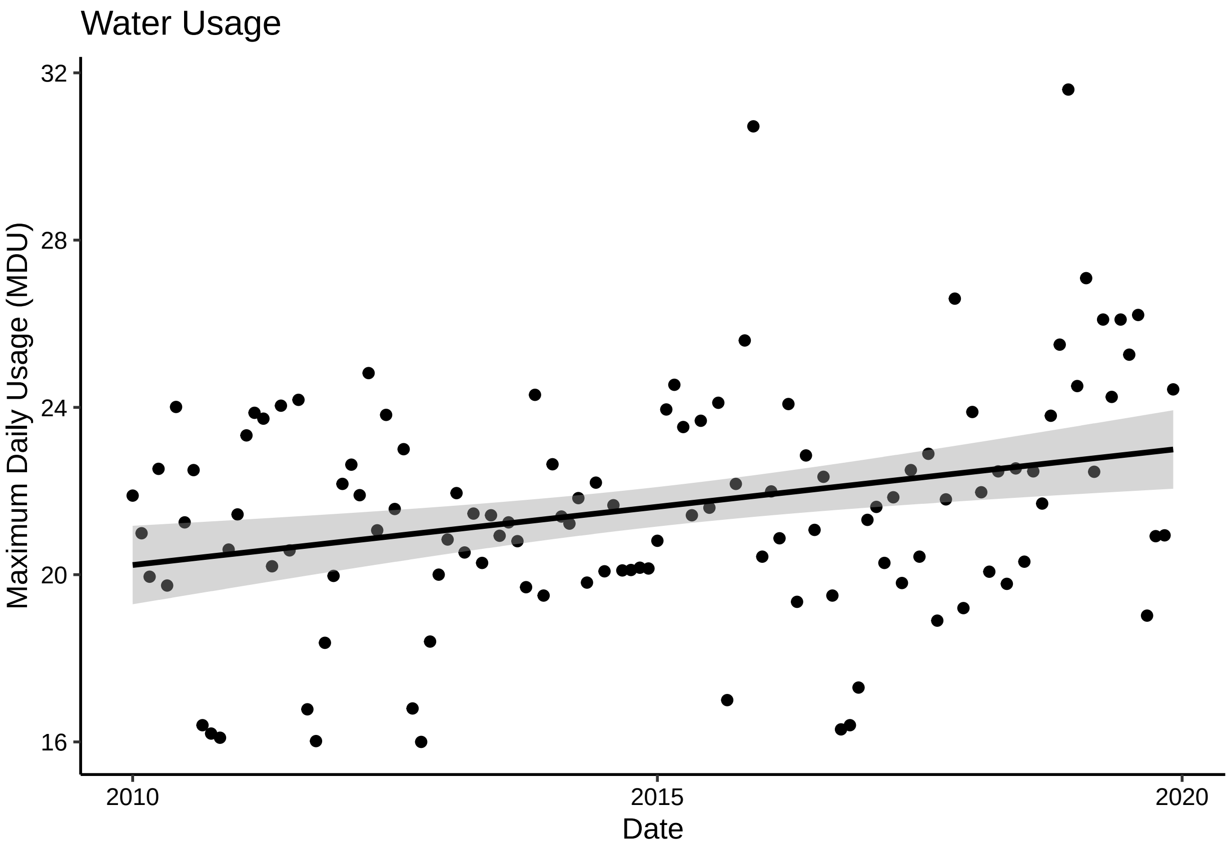
```
#8
scrape.it_Asheville <- function(the_year){
  #Get the proper url
  the_url <- ifelse(the_year==2021,
                    'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010',
                    paste('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=',
                          the_year,sep=''))

  #1 Link to the web site using read_html
  water_website <- read_html(the_url)
  #print(water_website)

  #2&3 Locate elements and read their text attributes into variables
  water.system.name <- water_website %>%
    html_nodes('td tr:nth-child(1) td:nth-child(2)') %>% html_text
  water.system.name <- water.system.name[1]
  pswid <- water_website %>%
    html_nodes('td tr:nth-child(1) td:nth-child(5)') %>% html_text
  ownership <- water_website %>%
    html_nodes('tr:nth-child(2) td:nth-child(4)') %>% html_text
  ownership <- ownership[1]
  max.withdrawals.mgd <- water_website %>%
    html_nodes(’th~ td:nth-child(6) , th~ td:nth-child(3)’) %>% html_text

  #3 Construct a dataframe from the values
  new_df <- data.frame(
  "Water System Name" = rep(water.system.name),
  "PSWID" = rep(pswid),
  "Ownership" = rep(ownership),
  "Maximum_Daily_Use" = max.withdrawals.mgd,
  month = rep(1:12),
  year = the_year

  )
  new_df <- new_df %>%
    mutate(Date = my(paste(month,"-",year)))
  return(new_df)
}

Asheville_2015_df <- scrape.it_Asheville(2015)
both_2015_df <-rbind(Durham_2015_df,Asheville_2015_df)

ggplot(both_2015_df, aes(x=Date,y=as.numeric(Maximum_Daily_Use), color = Water.System.Name)) +
  geom_point(size=3.0) +
  labs(title = "Water Usage",
       y="Maximum Daily Usage (MDU)",
       x="Date",
       color = "Location")
```



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
Asheville_2010_df <- scrape.it_Asheville(2010)
Asheville_2011_df <- scrape.it_Asheville(2011)
Asheville_2012_df <- scrape.it_Asheville(2012)
Asheville_2013_df <- scrape.it_Asheville(2013)
Asheville_2014_df <- scrape.it_Asheville(2014)
Asheville_2015_df <- scrape.it_Asheville(2015)
Asheville_2016_df <- scrape.it_Asheville(2016)
Asheville_2017_df <- scrape.it_Asheville(2017)
Asheville_2018_df <- scrape.it_Asheville(2018)
Asheville_2019_df <- scrape.it_Asheville(2019)


Asheville_df <- rbind(Asheville_2010_df, Asheville_2011_df, Asheville_2012_df,
                      Asheville_2013_df, Asheville_2014_df, Asheville_2015_df,
                      Asheville_2016_df, Asheville_2017_df, Asheville_2018_df,
                      Asheville_2019_df)

ggplot(Asheville_df, aes(x=Date,y=as.numeric(Maximum_Daily_Use)))+
  geom_point() +
  geom_smooth(method = lm, color = "black") +
  labs(title = "Water Usage",
       y="Maximum Daily Usage (MDU)",
       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?