# Assignment 09: Data Scraping

Halina Malinowski

## 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()
```

```
## [1] "C:/Users/Dell Laptop/Documents/GitHub/EDA/Assignments"
```

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

#The web address you put above was for 2020, but the directions said to use 2019.
#I went ahead and used the given URL for 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 <- NC_local_water_webpage %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)")%>%
  html_text()
water.system.name
```

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

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

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

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

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

```
max.withdrawals.mgd <- NC_local_water_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"
```

```
#The value given above as an example here is for 2020 and for the max day use...
#But the directions say to use Average Daily Use which is different it starts
#with 23.9700. I went ahead and used the max withdrawals (or Max Day Use) since
#that was the example and also the header for the code.
```

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...

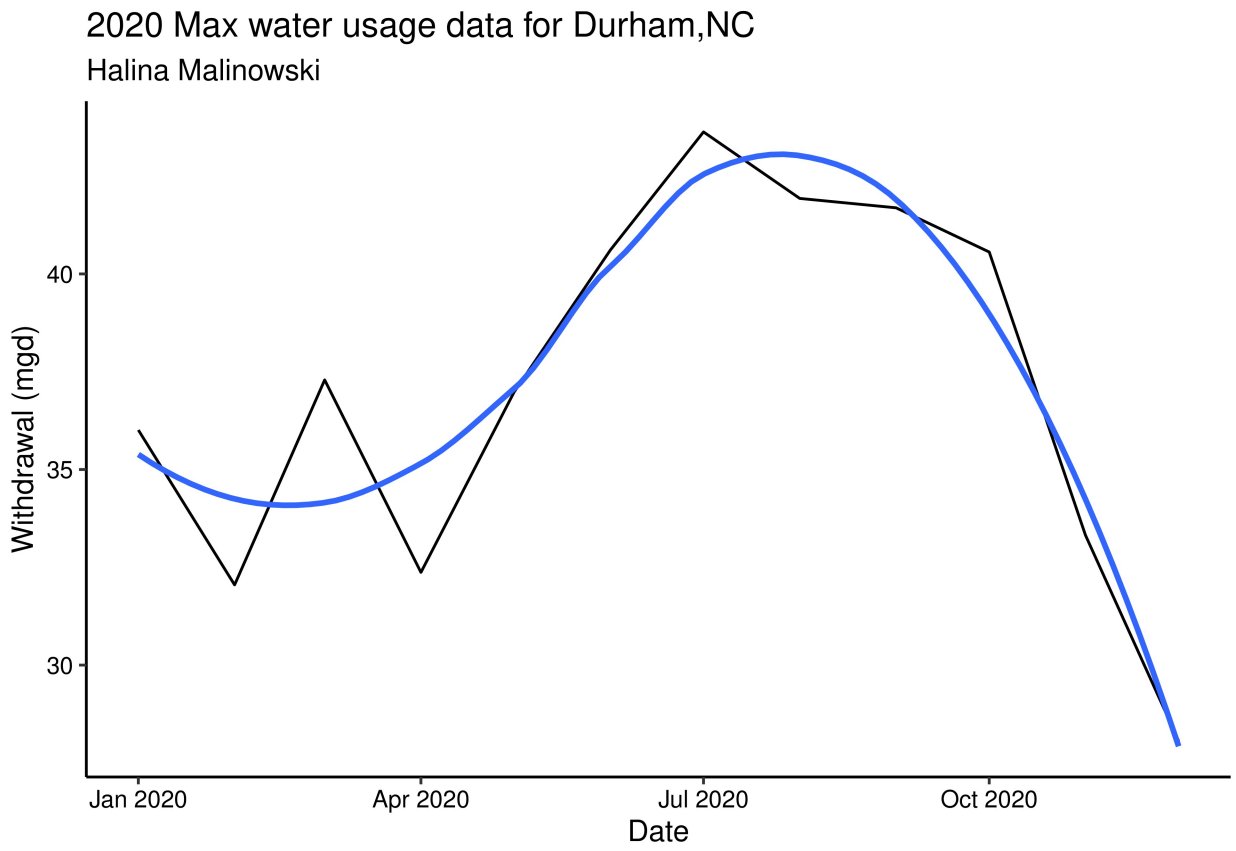
5. Plot the max daily withdrawals across the months for 2020

```
#4
NC_local_water_dataframe <- data.frame("Water_System_Name" = water.system.name,
           "Pwsid" = pwsid,
           "Ownership" = ownership,
           "Max_Withdrawals" = as.numeric(max.withdrawals.mgd),
           "Year" = rep(2020),
           "Month" = c("Jan", "May", "Sept", "Feb", "Jun", "Oct", "Mar",
                       "July", "Nov", "Apr", "Aug", "Dec"))

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

#5
ggplot(NC_local_water_dataframe,aes(x=Date,y= Max_Withdrawals)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = "2020 Max water usage data for Durham,NC",
       subtitle = "Halina Malinowski",
       y="Withdrawal (mgd)",
       x="Date")
```

```
## `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.
the_base_url <- 'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid='
the_pwsid <- '03-32-010'
the_year <- 2020
the_scrape_url <- paste0(the_base_url, the_pwsid, '&year=', the_year)
print(the_scrape_url)
```

```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2020"
```

```
scrape.it <- function(the_pwsid, the_year){

  #Retrieve the website contents
  the_website <- read_html(paste0(the_base_url, the_pwsid, '&year=', the_year))

  #Set the element tags
  water_system_name_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_tag <- 'th~ td+ td'

  #Scrape the data items
```

```
  water.system.name <- the_website %>% html_nodes(water_system_name_tag) %>%
    html_text()
  pwsid <- the_website %>%   html_nodes(pwsid_tag) %>%
    html_text()
  ownership <- the_website %>% html_nodes(ownership_tag) %>%
    html_text()
  max.withdrawals.mgd <- the_website %>% html_nodes(max_withdrawals_tag) %>%
    html_text()

  #Convert to a dataframe
  NC_local_water_dataframe <- data.frame("Water_System_Name" = water.system.name,
           "Pwsid" = pwsid,
           "Ownership" = ownership,
           "Max_Withdrawals" = as.numeric(max.withdrawals.mgd),
           "Year" = rep(the_year),
           "Month" = c("Jan", "May", "Sept", "Feb", "Jun", "Oct", "Mar",
                       "July", "Nov", "Apr", "Aug", "Dec"))

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

  #Return the dataframe
  return(NC_local_water_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
Durham_2015_Max_Withdrawals <- scrape.it('03-32-010', 2015)
view(Durham_2015_Max_Withdrawals)
```

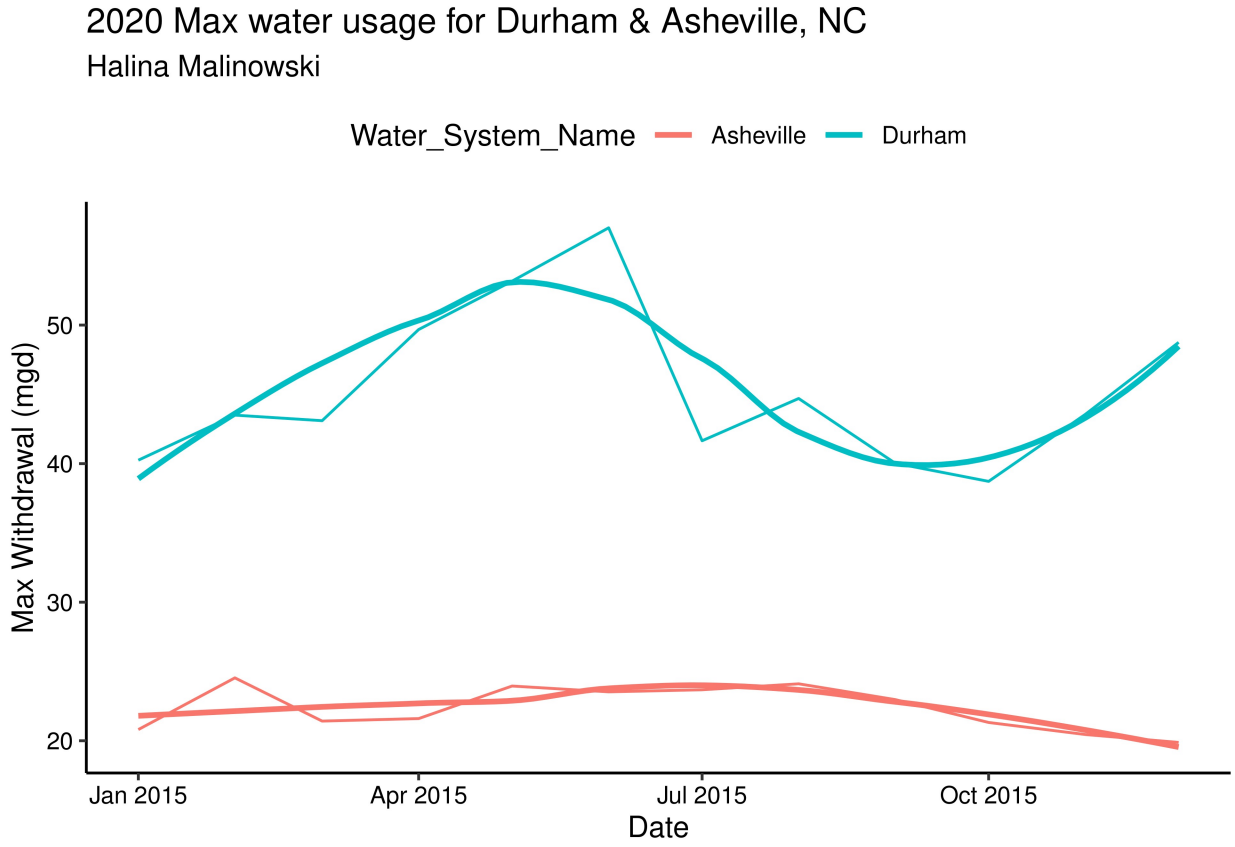
8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares the Asheville to Durham’s water withdrawals.

```
#8
Asheville_2015_Max_Withdrawals <- scrape.it('01-11-010', 2015)
view(Asheville_2015_Max_Withdrawals)

Durham_Asheville_2015 <- rbind(Durham_2015_Max_Withdrawals,
                               Asheville_2015_Max_Withdrawals)

ggplot(Durham_Asheville_2015,aes(x = Date,y = Max_Withdrawals)) +
  geom_line(aes(color = Water_System_Name)) +
  geom_smooth(aes(color = Water_System_Name), method="loess",se=FALSE, ) +
  labs(title = "2020 Max water usage for Durham & Asheville, NC",
       subtitle = "Halina Malinowski",
       y="Max Withdrawal (mgd)",
       x="Date")
```

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




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.

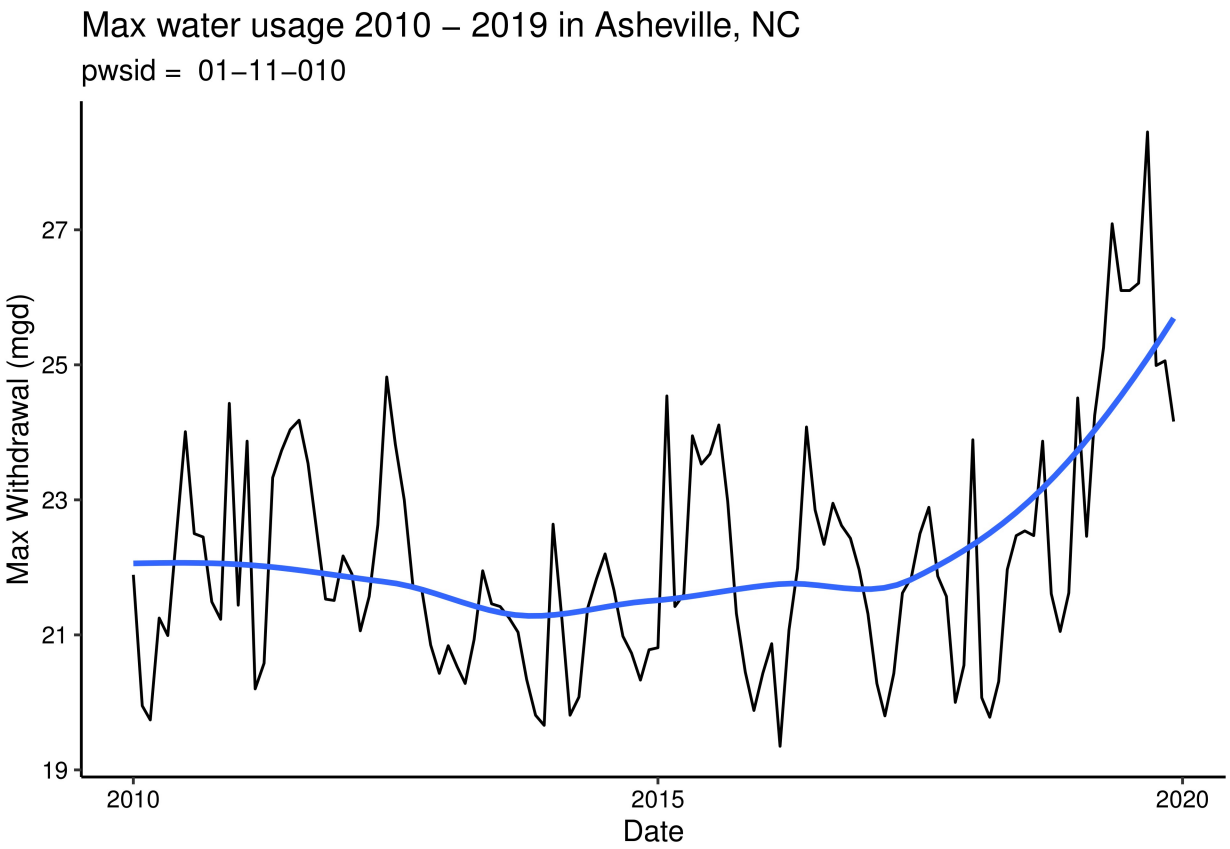
```
#9
the_years = rep(2010:2019)
my_pwsid = '01-11-010'

Asheville_Max_Withdrawals_2010_2019 <- lapply(X = the_years,
                  FUN = scrape.it,
                  the_pwsid = my_pwsid)

Combined_Asheville_2010_2019 <- bind_rows(Asheville_Max_Withdrawals_2010_2019)

ggplot(Combined_Asheville_2010_2019,aes(x = Date, y = Max_Withdrawals)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("Max water usage 2010 - 2019 in Asheville, NC"),
       subtitle = paste("pwsid = ", my_pwsid),
       y = "Max Withdrawal (mgd)",
       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?

Answer: Just by looking at the plot it appears that Asheville does have a trend in water usage over time. It appears that water usage has increased in Asheville from 2010 to 2020 with a steep increase in the last 3 years or so. A slight increase in max water use can start to be seen around 2015 after a dip or slight decrease around 2014. There may also be some seasonal trends in water usage as the data seem to oscillate between peaks and troughs. Perhaps this is due to a higher number of people during tourist seasons leading to greater water usage. Overall, Asheville's max water usage appears to be steadily increasing and further statistical analysis should be conducted to determine the significance of this increase, further trends, and possible causes, effects, and solutions.