

# Assignment 10: Data Scraping

Yixin Fang

## OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on data scraping.

## Directions

1. Rename this file `<FirstLast>_A10_DataScraping.Rmd` (replacing `<FirstLast>` with your first and last name).
2. Change “Student Name” on line 3 (above) with your name.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure your code is tidy; use line breaks to ensure your code fits in the knitted output.
5. Be sure to **answer the questions** in this assignment document.
6. When you have completed the assignment, **Knit** the text and code into a single PDF file.

## Set up

1. Set up your session:
  - Load the packages `tidyverse`, `rvest`, and any others you end up using.
  - Check your working directory

```
#1
Sys.setenv("LANGUAGE"="En")
Sys.setlocale("LC_ALL", "en_GB.UTF-8")
```

```
## [1] "LC_COLLATE=en_GB.UTF-8;LC_CTYPE=en_GB.UTF-8;LC_MONETARY=en_GB.UTF-8;LC_NUMERIC=C;LC_TIME=en_GB.UTF-8"
```

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

```
## [1] "D:/DKU/2023_Spring/ENV872/EDA-Spring2023"
```

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham’s 2022 Municipal Local Water Supply Plan (LWSP):
  - Navigate to <https://www.ncwater.org/WUDC/app/LWSP/search.php>
  - 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=2022>

Indicate this website as the as the URL to be scraped. (In other words, read the contents into an `rvest` webpage object.)

```
#2
website <- read_html("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2022")
```

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:
- Maximum Day Use (MGD) - for each month

In the code chunk below scrape these values, assigning them to four 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 (represented as strings), with the first value being “36.1000”.

```
#3
water.system.name <- website %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()

PWSID <- website %>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
  html_text()

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

max.withdrawals.mgd <- website %>%
  html_nodes("th~ td+ td") %>%
  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.)

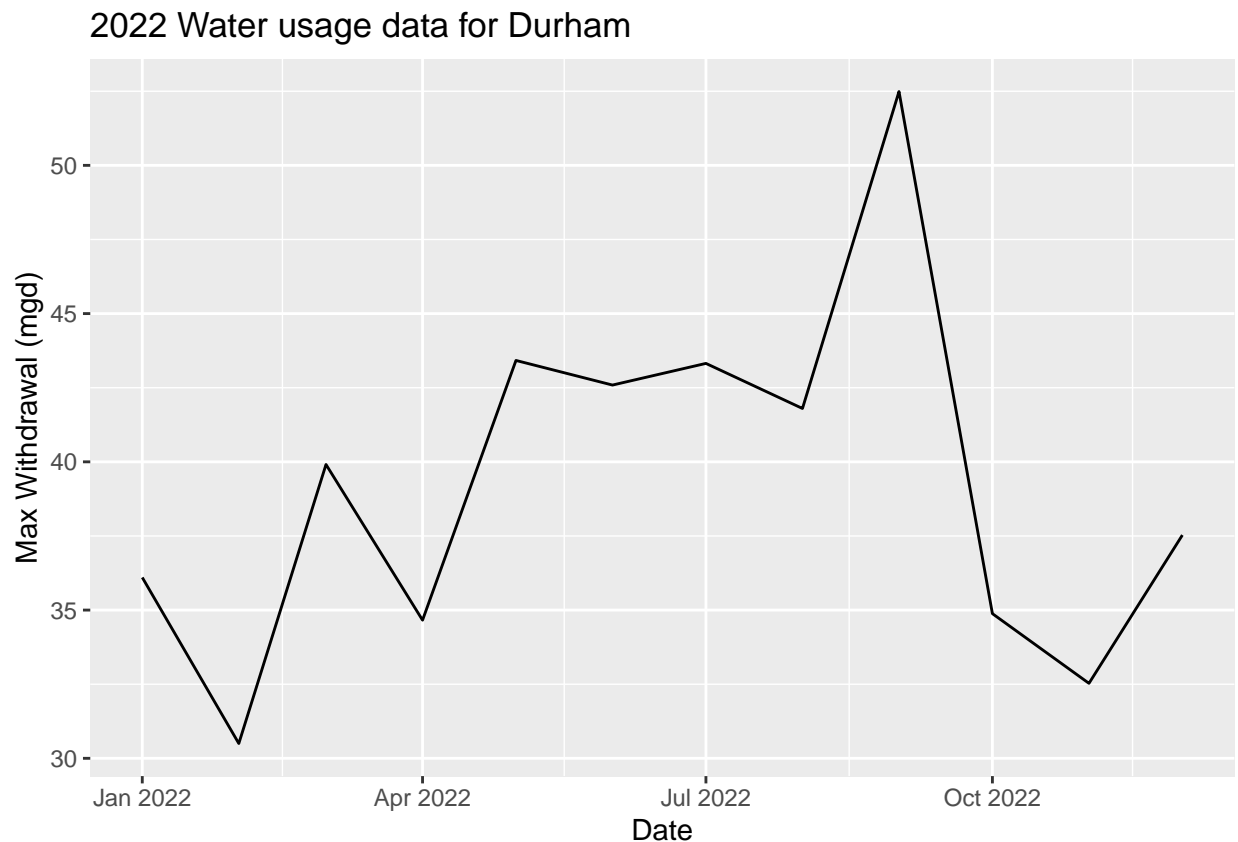
TIP: Use `rep()` to repeat a value when creating a dataframe.

NOTE: It’s likely you won’t be able to scrape the monthly withdrawal data in chronological order. You can overcome this by creating a month column manually assigning values in the order the data are scraped: “Jan”, “May”, “Sept”, “Feb”, etc... Or, you could scrape month values from the web page...

5. Create a line plot of the max daily withdrawals across the months for 2022

```
#4
df_mgd <- data.frame("Water_System_name" = rep(water.system.name,12),
                     "Ownership" = rep(ownership,12),
                     "Max-Withdrawals_mgd" = as.numeric(max.withdrawals.mgd),
                     "PWSID" = rep(PWSID,12),
                     "Month" = c(1,5,9,2,6,10,3,7,11,4,8,12),
                     "Year" = rep(2022,12)) %>%
  mutate(Date = as.Date(paste(Year,Month,"01", sep = "-")))

#5
ggplot(df_mgd,aes(x=Date,y=Max-Withdrawals_mgd)) +
  geom_line() +
  labs(title = paste("2022 Water usage data for Durham"),
       y="Max Withdrawal (mgd)",
       x="Date")
```



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 (pwsid) scraped.**

```

#6.
scrape.it <- function(the_year, the_PWSID){

  the_website <- read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',
                                   the_PWSID, '&year=', the_year))
  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.mgd_tag <- "th~ td+ td"

  #Scrape the data items
  the.water.system.name <- the_website %>% html_nodes(water.system.name_tag) %>% html_text()
  the.PWSID <- the_website %>% html_nodes(PWSID_tag) %>% html_text()
  the.ownership <- the_website %>% html_nodes(ownership_tag) %>% html_text()
  the.withdrawals.mgd <- the_website %>% html_nodes(max.withdrawals.mgd_tag) %>% html_text()

  #Convert to a dataframe
  df_withdrawals <- data.frame("Water_System_name" = rep(the.water.system.name,12),
                              "Ownership" = rep(the.ownership,12),
                              "Max-Withdrawals_mgd" = as.numeric(the.withdrawals.mgd),
                              "PWSID" = rep(the.PWSID,12),
                              "Month" = c(1,5,9,2,6,10,3,7,11,4,8,12),
                              "Year" = rep(the_year,12)) %>%
  mutate(Date = as.Date(paste(Year,Month,"01", sep = "-")))

  return(df_withdrawals)
}

```

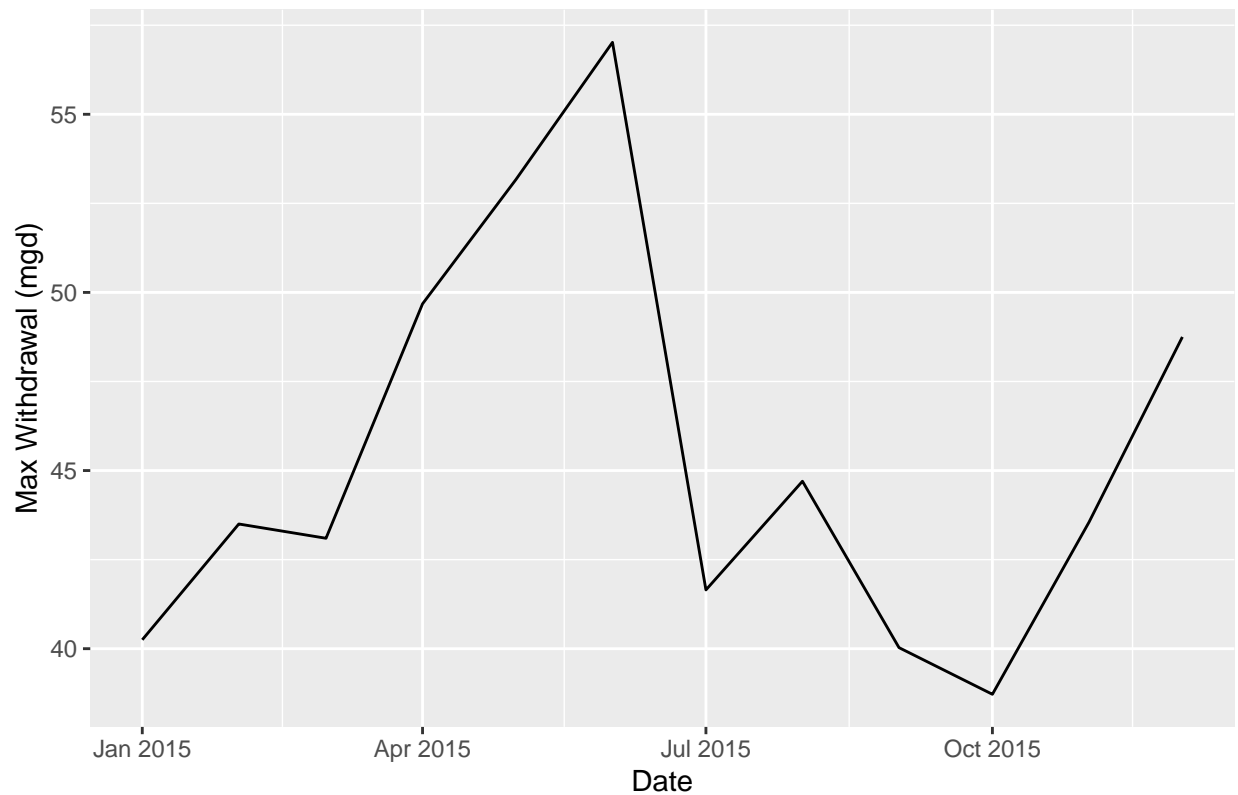
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID='03-32-010') for each month in 2015

```

#7
df2015 <- scrape.it(2015,"03-32-010")
ggplot(df2015,aes(x=Date,y=Max-Withdrawals_mgd)) +
  geom_line() +
  labs(title = paste("2015 Water usage data for Durham"),
       y="Max Withdrawal (mgd)",
       x="Date")

```

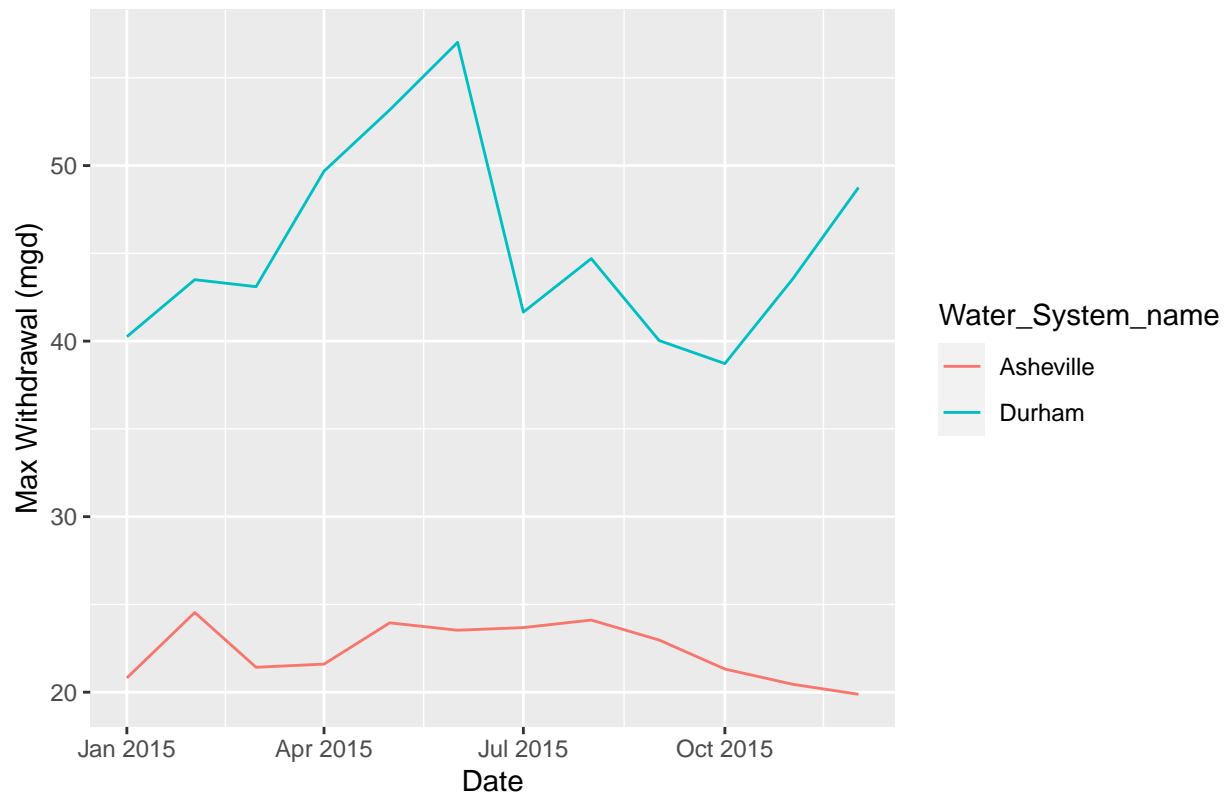
## 2015 Water usage data for Durham



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 Asheville's to Durham's water withdrawals.

```
#8
dfasheville <- scrape.it(2015,"01-11-010")
join <- rbind(dfasheville,df2015)
ggplot(join,aes(x=Date,y=Max-Withdrawals_mgd)) +
  geom_line(aes(color = Water_System_name)) +
  labs(title = paste("2015 Water usage data for Durham and Asheville"),
       y="Max Withdrawal (mgd)",
       x="Date")
```

## 2015 Water usage data for Durham and Asheville



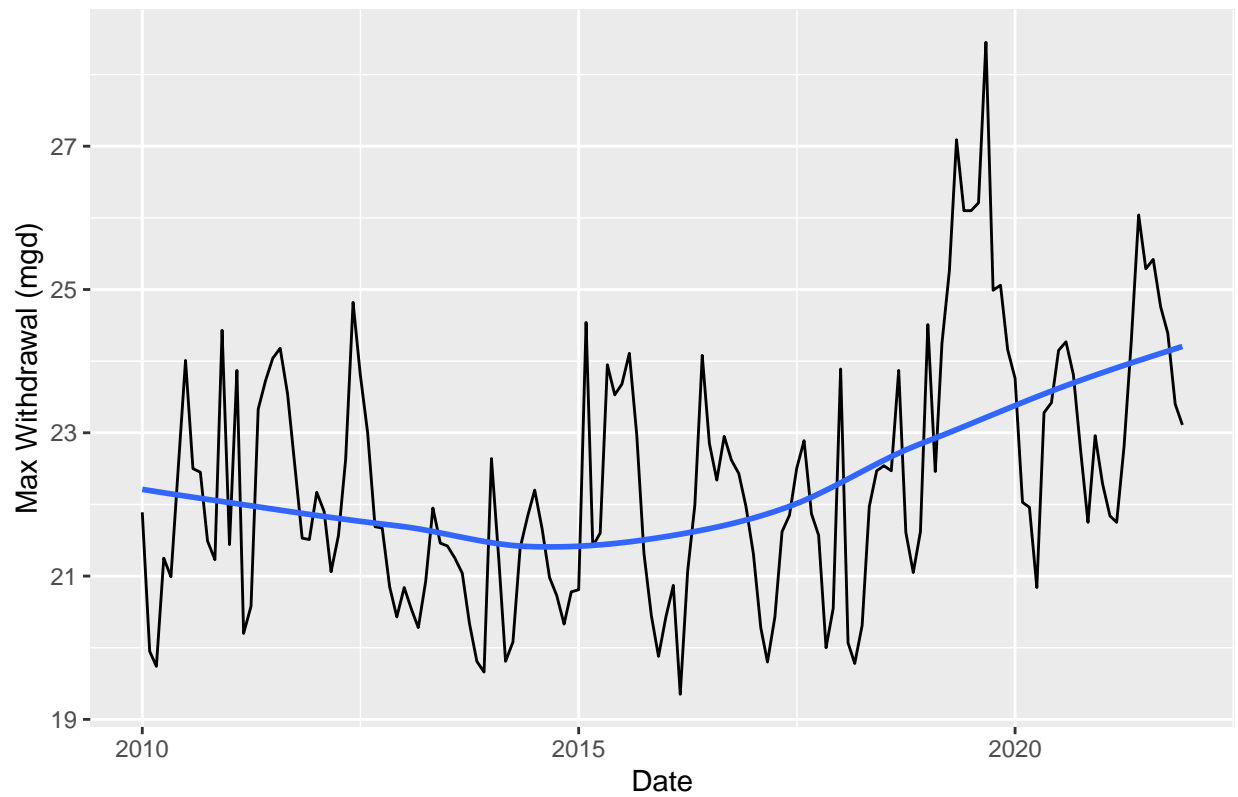
- Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2021. Add a smoothed line to the plot (method = 'loess').

TIP: See Section 3.2 in the "09\_Data\_Scraping.Rmd" where we apply "map2()" to iteratively run a function over two inputs. Pipe the output of the map2() function to `bind_rows()` to combine the dataframes into a single one.

```
#9
yearlist <- rep(2010:2021)
PWSIDlist <- rep("01-11-010",12)
dfs_asheville <- map2(yearlist, PWSIDlist, scrape.it) %>%
  bind_rows()
ggplot(dfs_asheville, aes(x=Date, y=Max-Withdrawals_mgd)) +
  geom_line() +
  geom_smooth(method="loess", se=FALSE) +
  labs(title = paste("2010-2021 Water usage data for Asheville"),
       y="Max Withdrawal (mgd)",
       x="Date")
```

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

2010–2021 Water usage data for Asheville



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

Answer: Yes. It looks like there is an increasing trend in the water usage.