

Assignment 09: Data Scraping

Laura Martinez

OVERVIEW

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

Directions

1. Rename this file `<FirstLast>_A09_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 to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.

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
setwd("~/Documents/EDA-Fall2022")
getwd()

## [1] "/Users/laura/Documents/EDA-Fall2022"

#install.packages(tidyverse)
library(tidyverse)
#install.packages(rvest)
library(rvest)
#install.packages(lubridate)
library(lubridate)
library(ggplot2)

#Set theme
Laurastheme <- theme_classic(base_size = 12) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "right")
theme_set(Laurastheme)
```

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham’s 2021 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=2021>

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

```
#2
theURL <- 'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2021'
the_main_URL <- read_html(theURL)
```

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 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 “27.6400”.

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

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

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

max.withdrawals.mgd <- the_main_URL %>%
  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. . .

5. Create a line plot of the maximum daily withdrawals across the months for 2021

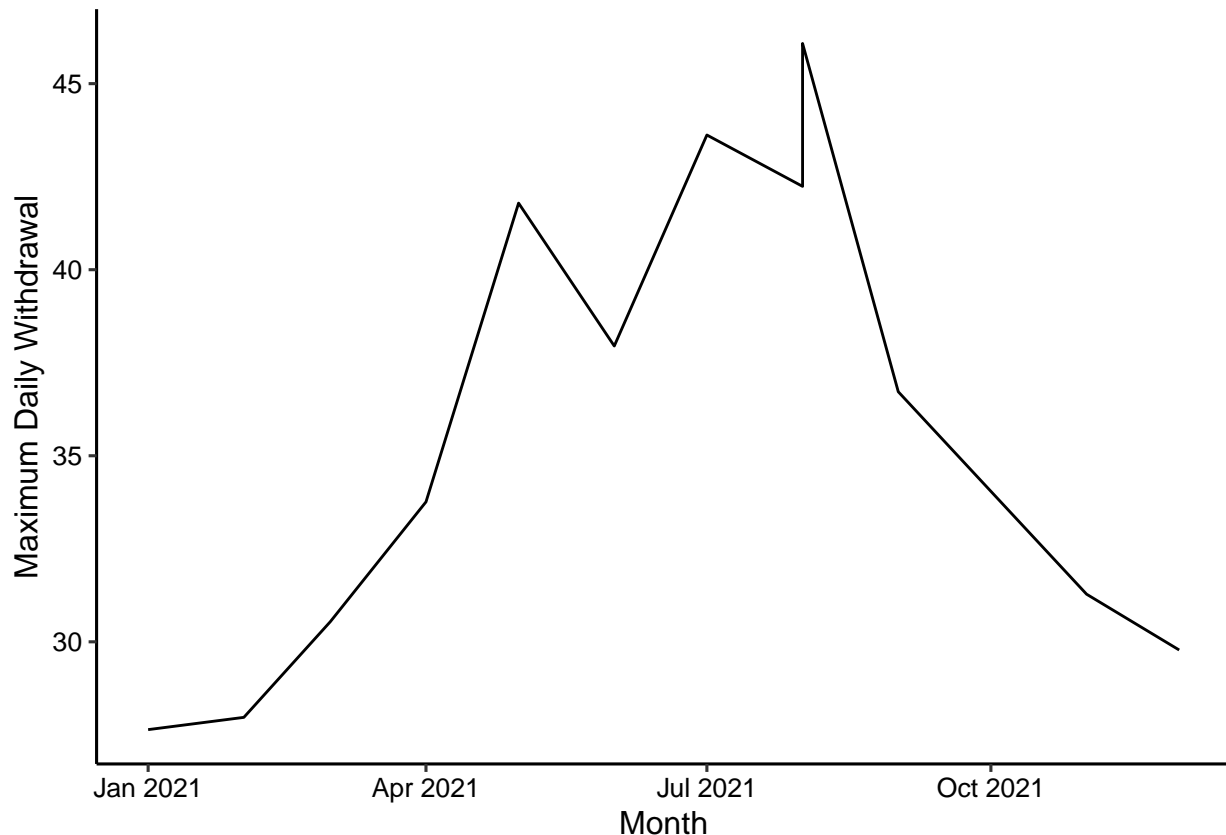
```
#4
df <- data.frame("Month" = c(1, 5, 9, 2, 6, 8, 3, 7, 11, 4, 8, 12),
  "Year" = rep(2021,12),
  "Water System Name" = water.system.name,
  "PSWID" = pwsid,
  "Ownership" = ownership,
```

```

      "Max-Withdrawals_mgd" = as.numeric(max.withdrawals.mgd)) %>%
mutate(Date = my(paste(Month, "-", Year)))

#5
max.plot <- ggplot(data = df, aes(x = Date, y = Max-Withdrawals_mgd)) +
  geom_line() +
  xlab("Month") +
  ylab("Maximum Daily Withdrawal")
max.plot

```



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_pwsid, the_year){

  the_scrape_url <- read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?',
    'pwsid=', the_pwsid, '&', 'year=', the_year))

  #Scrape the data items
  water.system.name <- the_scrape_url %>%
    html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>% html_text()
  the_pwsid <- the_scrape_url %>%
    html_nodes("td tr:nth-child(1) td:nth-child(5)") %>% html_text()
  ownership <- the_scrape_url %>%
    html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>% html_text()
  max.withdrawals.mgd <- the_scrape_url %>% html_nodes("th~ td+ td") %>% html_text()
}

```

```

#Construct dataframe
dataframe <- data.frame("Month" = c(1, 5, 9, 2, 6, 8, 3, 7, 11, 4, 8, 12),
                        "Year" = rep(the_year,12),
                        "Water System Name" = water.system.name,
                        "PSWID" = the_pswid,
                        "Ownership" = ownership,
                        "Max-Withdrawals_mgd" = as.numeric(max.withdrawals.mgd)) %>%
  mutate(Date = my(paste(Month,"-",Year)))
return(dataframe)
}

```

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

```

#7
df_Dur2015 <- scrape.it('03-32-010', 2015)

max.plot_Dur2015 <- ggplot(data = df_Dur2015, aes(x = Date, y = Max-Withdrawals_mgd)) +
  geom_line() +
  xlab("Month") +
  ylab("Maximum Daily Withdrawal")
max.plot_Dur2015

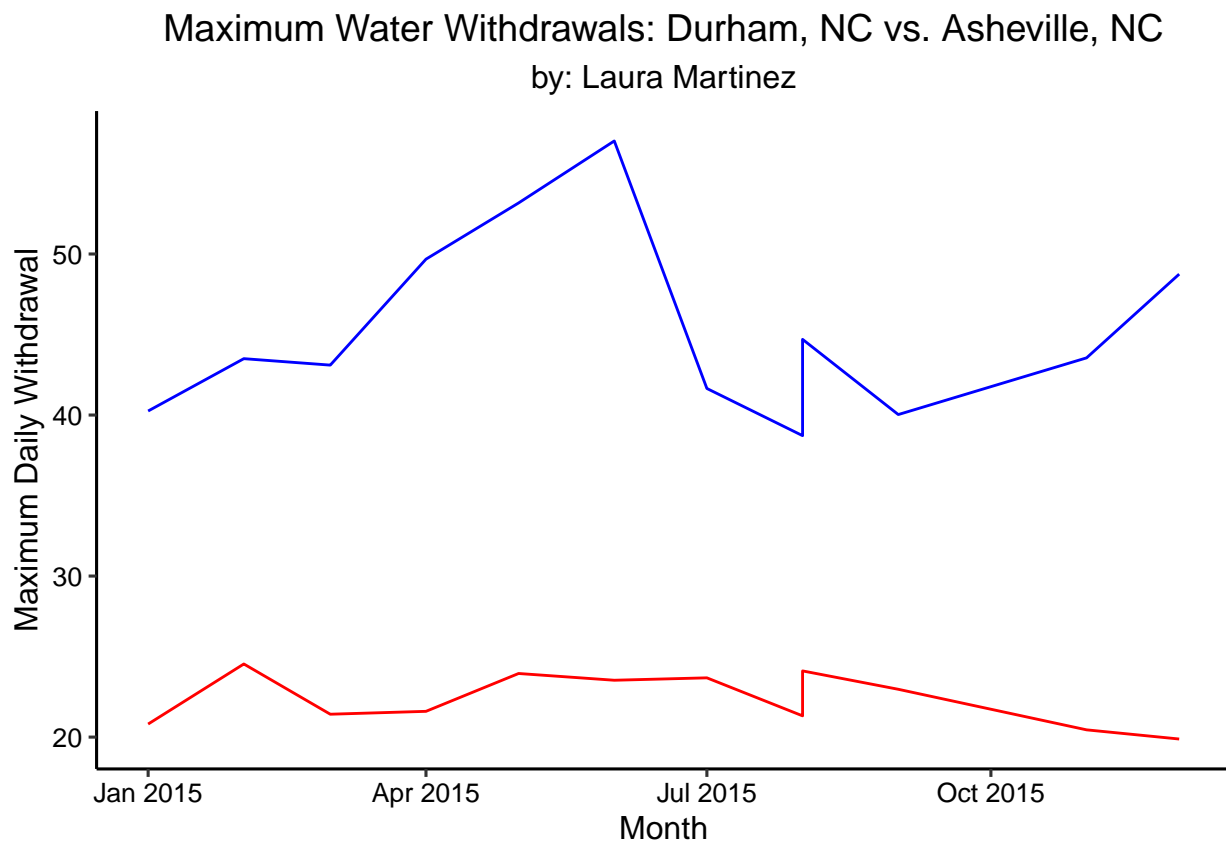
```



- 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
df_Ash2015 <- scrape.it('01-11-010', 2015)

max.plot_2015 <- ggplot() +
  geom_line(data = df_Dur2015, aes(x = Date, y = Max-Withdrawals_mgd), color = "blue" ) +
  geom_line(data = df_Ash2015, aes(x = Date, y = Max-Withdrawals_mgd), color = "red" ) +
  xlab("Month") +
  ylab("Maximum Daily Withdrawal") +
  labs(title = "Maximum Water Withdrawals: Durham, NC vs. Asheville, NC",
       subtitle = "by: Laura Martinez") +
  theme(plot.title=element_text(hjust=0.5),
        plot.subtitle=element_text(hjust=0.5))
max.plot_2015
```



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.

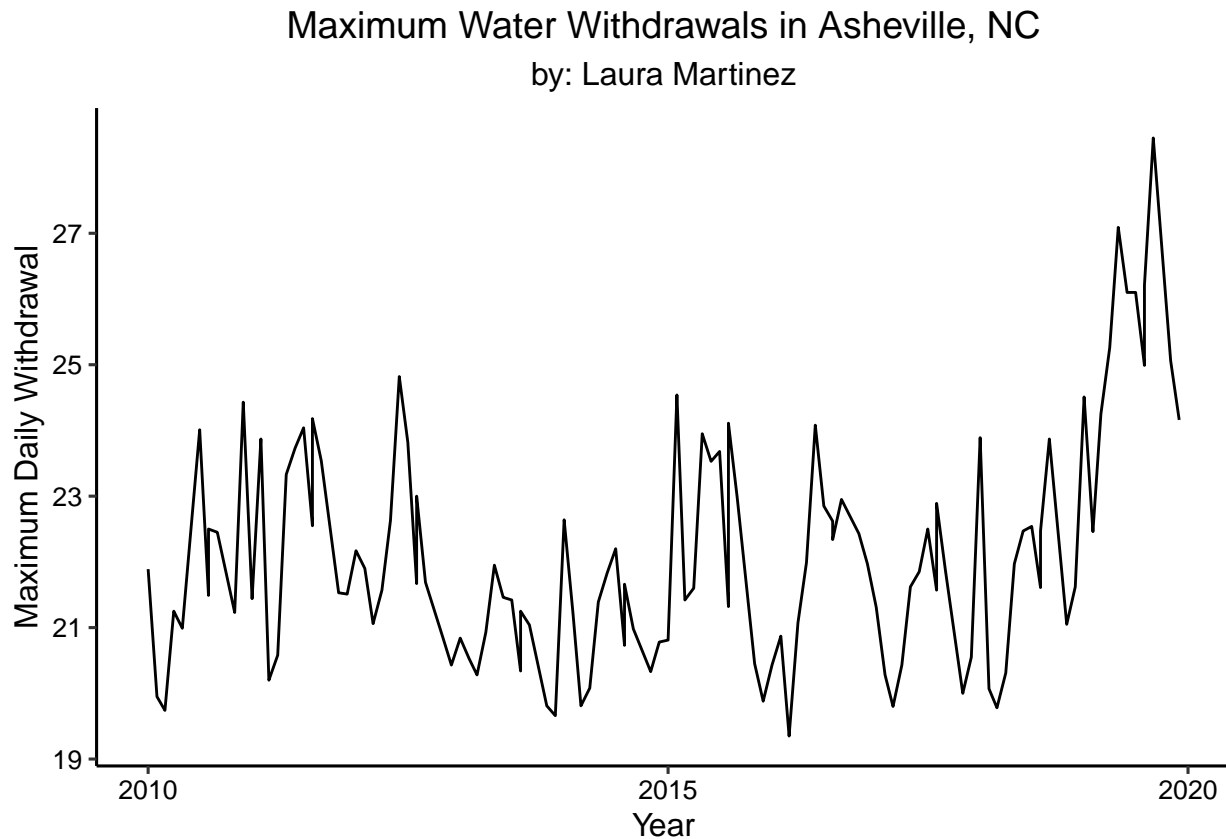
TIP: See Section 3.2 in the "09_Data_Scraping.Rmd" where we apply "map2()" to iteratively run a function over two inputs. Pipe the output of the map2() function to bind_rows() to combine the dataframes into a single one.

```
#9
year_range <- seq(2010, 2019)

Asheville_pswid <- rep('01-11-010', length(year_range))

Asheville_df <- map2(Asheville_pswid, year_range, scrape.it) %>% bind_rows()
```

```
Asheville.plot <- ggplot(data = Asheville_df, aes(x = Date, y = Max-Withdrawals_mgd)) +
  geom_line() +
  xlab("Year") +
  ylab("Maximum Daily Withdrawal") +
  labs(title = "Maximum Water Withdrawals in Asheville, NC",
       subtitle = "by: Laura Martinez") +
  theme(plot.title=element_text(hjust=0.5),
       plot.subtitle=element_text(hjust=0.5))
Asheville.plot
```



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

Asheville's water consumption has a cyclical nature where there are observed fluctuations in water usage that occur periodically. In general, the water levels seem to be consistent over the years. That is until the dramatic spike in late 2019/2020, which suggests that there may be something happening externally other than seasonal trends that could explain an uptick in water consumption in Asheville, NC during this time.