

# Assignment 09: Data Scraping

Nicole Eastman

## 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 Set-up
library(formatR)
library(tidyr)
knitr::opts_chunk$set(tidy.opts=list(width.cutoff=60), tidy=TRUE)

getwd()

## [1] "/home/guest/R/EDA Fall 2022"

library(tidyverse)
library(rvest)
library(lubridate)
library(ggplot2)
library(dataRetrieval)

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 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?psid=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 Using rvest to scrape website*

```
theURL <- read_html(paste0("https://www.ncwater.org/WUDC/app/LWSP/report.php?psid=03-32-010&year=2021"))
```

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 Use the Selector Gadget tool to assign values to*

*# variables*

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

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

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

```
max.withdrawals.mgd <- theURL %>%
  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 Create dataframe with variables, convert month to order*

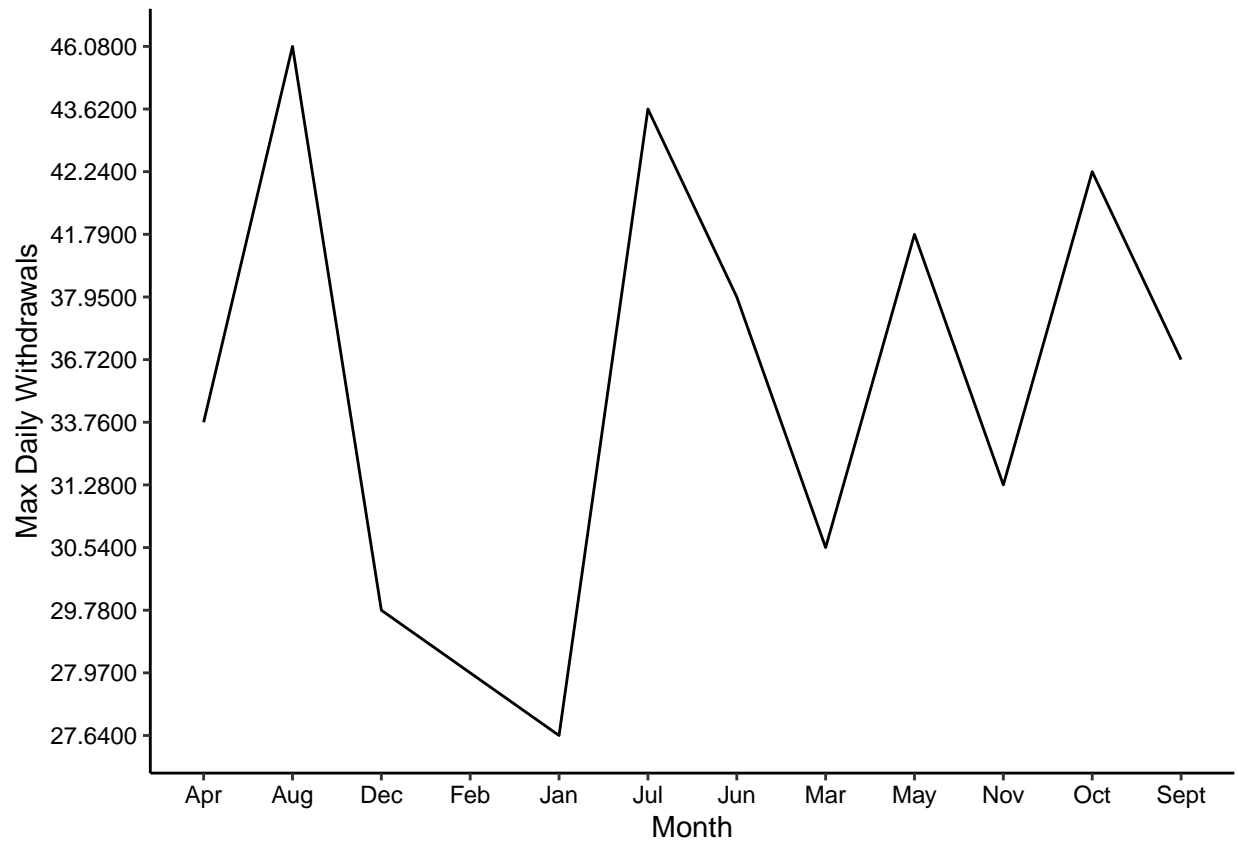
*# they are scraped from the website, and mutate to create a*

```
# date format column
Month <- c("01", "02", "03", "04", "05", "06", "07", "08", "09",
          "10", "11", "12")

df_withdrawals <- data.frame(Water_System = water.system.name,
                             PSWID = pswid, Ownership = ownership, Max-Withdrawals = max.withdrawals.mgd,
                             Month = rep(1:12), Year = rep(2021))
df_withdrawals$Month <- c("Jan", "May", "Sept", "Feb", "Jun",
                          "Oct", "Mar", "Jul", "Nov", "Apr", "Aug", "Dec")

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

# 5
ggplot(df_withdrawals, aes(x = Month, y = Max-Withdrawals, group = 1)) +
  geom_line() + labs(x = "Month", y = "Max Daily Withdrawals")
```



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 (pswid) scraped.

```
# 6.
scrape.it <- function(the_pswid, Year) {
  the_website <- read_html(paste0("https://www.ncwater.org/WUDC/app/LWSP/report.php?pswid=",
    the_pswid, "&year=", Year))
  water.system.name <- the_website %>%
    html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
```

```

    html_text()

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

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

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

    df_PSWID <- data.frame(Month = (Month), Year = Year, Max.withdrawals = as.numeric(max.withdrawals.mgd),
      Date = as.Date(paste0(Month, "-", Year), format = "%M-%Y")) %>%
      mutate(Ownership = !!ownership, PWSID = !!pswid, System.name = !!water.system.name)
    return(df_PSWID)
}

```

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_PSWID1 <- scrape.it("03-32-010", 2015)

```

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.

```

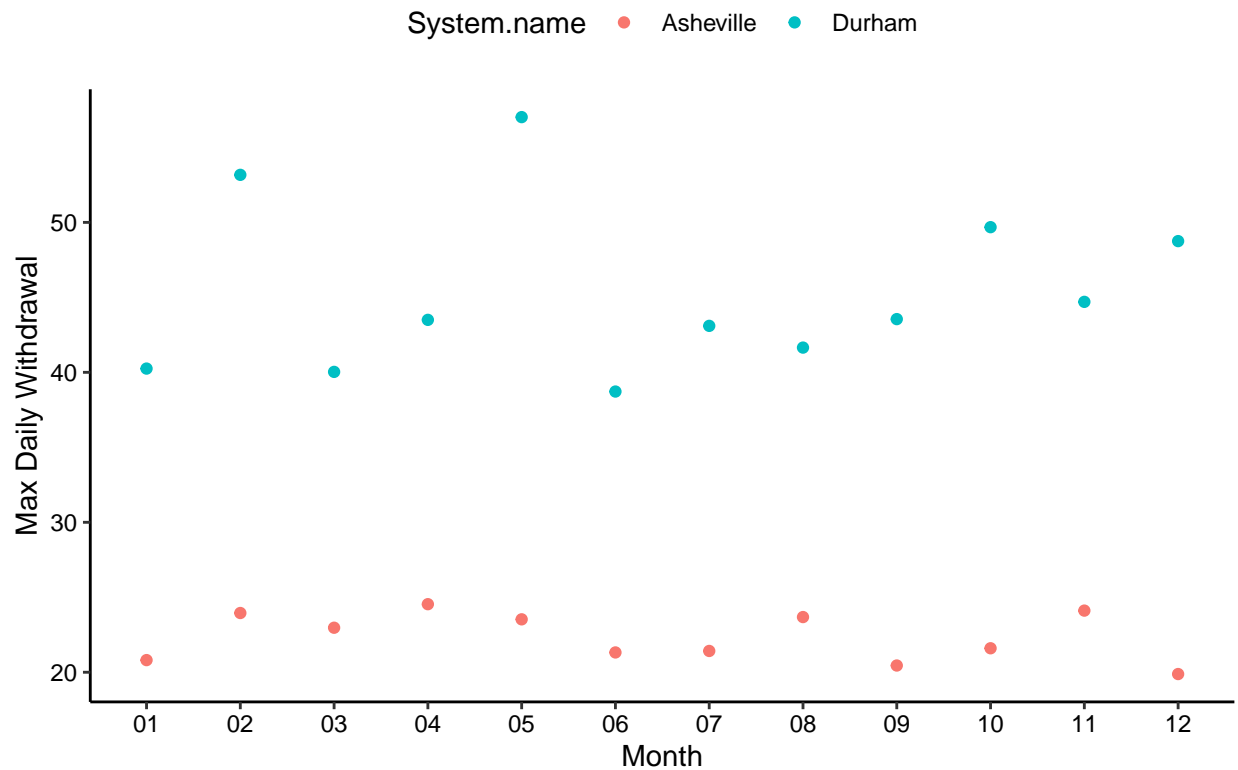
# 8
df_Ashville2015 <- scrape.it("01-11-010", 2015)

joined_df <- rbind(df_Ashville2015, df2015_PSWID1)

combined.plot <- ggplot(joined_df, aes(x = Month, y = Max.withdrawals,
  color = System.name)) + geom_point() + labs(title = "Durham and Asheville Maximum Daily Water Withdrawals",
  y = "Max Daily Withdrawal", x = "Month")
combined.plot

```

## Durham and Asheville Maximum Daily Water Withdrawals



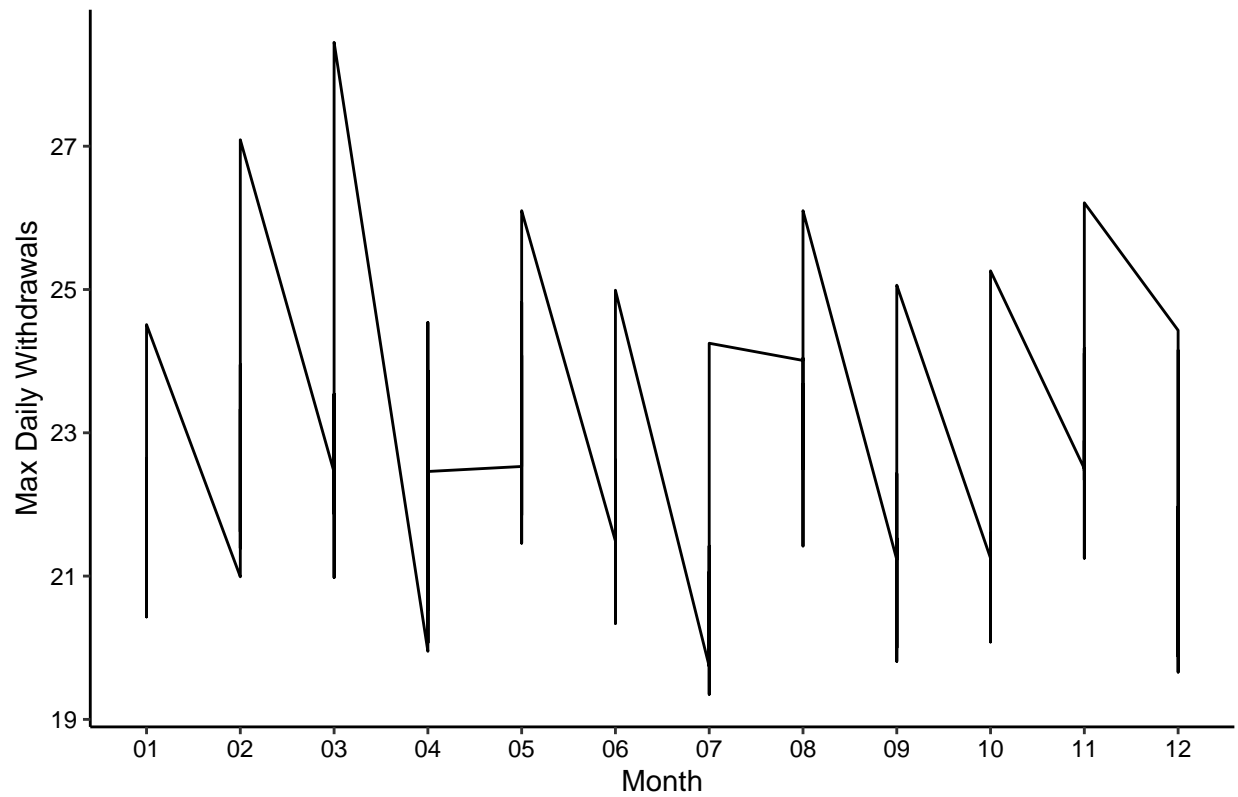
9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019. Add a smoothed line to the plot.

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
the_Years <- rep(2010:2019)
the_PWSID <- rep("01-11-010", length(the_Years))
df_many_years <- map2(the_PWSID, the_Years, scrape.it) %>%
  bind_rows()

asheville.plot <- ggplot(df_many_years, aes(x = Month, y = Max.withdrawals,
  group = 2)) + geom_line() + labs(x = "Month", y = "Max Daily Withdrawals",
  title = "Max Daily Withdrawal from 2010-2019 in Asheville")
asheville.plot
```

Max Daily Withdrawal from 2010–2019 in Asheville



Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time? The graph indicates there may be a slight upward trend in Asheville's water usage over time.