Assignment 10: Data Scraping

Sam Campbell

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
#load packages
library(tidyverse)
library(rvest)
library(lubridate)
library(viridis)
library(here)

install.packages("rvest")
library(rvest)

install.packages("dataRetrieval")
library(dataRetrieval)

install.packages("tidycensus")
library(tidycensus)

#set theme
mytheme <- theme_classic() +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
```

```
theme_set(mytheme)

#check working directory
getwd()
```

```
## [1] "/home/guest/R/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 URL to be scraped. (In other words, read the contents into an `rvest` webpage object.)

```
#2
#indicate website as the URL to be scraped
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
#Scrape Water System Name
water.system.name <- html_nodes(website,
                                'div+ table tr:nth-child(1) td:nth-child(2)') %>%
  html_text()

#Scrape PWSID
PWSID <- html_nodes(website, 'td tr:nth-child(1) td:nth-child(5)') %>%
  html_text()
```

```
#Scrape Ownership
ownership <- html_nodes(website, 'div+ table tr:nth-child(2) td:nth-child(4)') %>%
  html_text()

#Scrape Maximum Day Use for each month
max.withdrawals.mgd <- html_nodes(website, 'th~ td+ td') %>%
  html_text()
max.withdrawals.mgd
```

```
##  [1] "36.1000" "43.4200" "52.4900" "30.5000" "42.5900" "34.8800" "39.9100"
##  [8] "43.3200" "32.5300" "34.6600" "41.8000" "37.5300"
```

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 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
#Build dates vector
#Date <- c("2022-01", "2022-05", "2022-09", "2022-02", "2022-06", "2022-10",
            "2022-03", "2022-07", "2022-11", "2022-04", "2022-08", "2022-12")

#Convert dates vector to date format
#Date <- as.Date(Date)

#Check class of Date vector
#class(Date)

#Convert scraped data into data frame
#ncwaterdf <- data.frame(WaterSystemName = water.system.name,
                         #PWSID = PWSID,
                         #Ownership = ownership,
                         #MaxDayUse = max.withdrawals.mgd,
                         #Month = as_date(c("Jan", "May", "Sep", "Feb", "Jun", "Oct",
                                   "Mar", "Jul", "Nov", "Apr", "Aug", "Dec")),
                         #Date = Date,
                         #Year = rep(2022))

#2nd try at data frame, adjusting month approach
ncwaterdf <- data.frame(WaterSystemName = water.system.name,
                         PWSID = PWSID,
                         Ownership = ownership,
                         MaxDayUse = max.withdrawals.mgd,
```

```
                        Month = month(c(1,5,9,2,6,10,3,7,11,4,8,12)),
                        Year = rep(2022))

#Change class of Month column to a date
as_date(ncwaterdf$Month)

#5
#Line plot of max daily withdrawals across months
MonthlyWithdrawalsPlot <-
  ggplot(ncwaterdf,
       aes(x = Month,
           y = MaxDayUse)) +
  geom_point() +
  geom_line() +
  labs(y="Withdrawal (mgd)",
       x="Date")
print(MonthlyWithdrawalsPlot)
```

```
## Error: <text>:4:22: unexpected ','
## 3: #Date <- c("2022-01", "2022-05", "2022-09", "2022-02", "2022-06", "2022-10",
## 4:             "2022-03",
##                         ^
```

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.
#Set the URL to be scraped, modifying PWSID and Year
website2 <- read_html('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2015')

#Scrape Water System Name
water.system.name2 <- html_nodes(website2,
                                 'div+ table tr:nth-child(1) td:nth-child(2)') %>%
  html_text()

#Scrape PWSID
PWSID2 <- html_nodes(website2, 'td tr:nth-child(1) td:nth-child(5)') %>%
  html_text()

#Scrape Ownership
ownership2 <- html_nodes(website2, 'div+ table tr:nth-child(2) td:nth-child(4)') %>%
  html_text()

#Scrape Maximum Day Use for each month
max.withdrawals.mgd2 <- html_nodes(website2, 'th~ td+ td') %>%
  html_text()
max.withdrawals.mgd
```

```
##  [1] "36.1000" "43.4200" "52.4900" "30.5000" "42.5900" "34.8800" "39.9100"
##  [8] "43.3200" "32.5300" "34.6600" "41.8000" "37.5300"
```

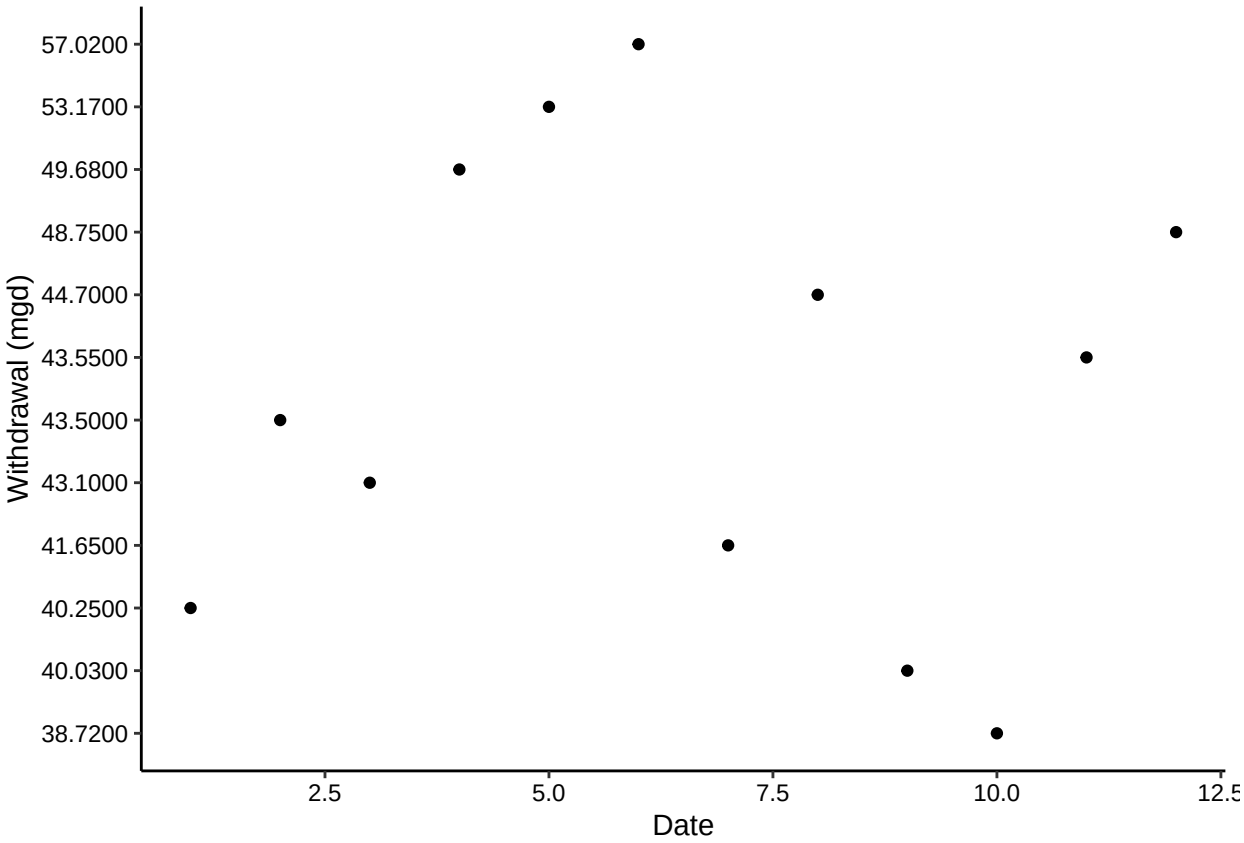
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID=‘03-32-010’) for each month in 2015

```
#7
#Set the URL to be scraped, modifying PWSID and Year
website2 <- read_html('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2015')

#Scrape Water System Name
water.system.name2 <- html_nodes(website2,
                                 'div+ table tr:nth-child(1) td:nth-child(2)') %>%
  html_text()

#Scrape PWSID
PWSID2 <- html_nodes(website2, 'td tr:nth-child(1) td:nth-child(5)') %>%
  html_text()

#Scrape Ownership
ownership2 <- html_nodes(website2, 'div+ table tr:nth-child(2) td:nth-child(4)') %>%
  html_text()

#Scrape Maximum Day Use for each month
max.withdrawals.mgd2 <- html_nodes(website2, 'th~ td+ td') %>%
  html_text()
max.withdrawals.mgd
```

```
##  [1] "36.1000" "43.4200" "52.4900" "30.5000" "42.5900" "34.8800" "39.9100"
##  [8] "43.3200" "32.5300" "34.6600" "41.8000" "37.5300"
```

```
#Construct data frame
ncwaterdf.2015 <- data.frame(WaterSystemName = water.system.name2,
                         PWSID = PWSID2,
                         Ownership = ownership2,
                         MaxDayUse = max.withdrawals.mgd2,
                         Month = month(c(1,5,9,2,6,10,3,7,11,4,8,12)),
                         Year = rep(2015))

#Plot data
Plot2015 <-
  ggplot(ncwaterdf.2015, aes(x = Month, y = MaxDayUse)) +
  geom_point() +
  geom_line() +
  labs(y="Withdrawal (mgd)",
       x="Date")
print(Plot2015)
```

```
## ‘geom_line()‘: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
```

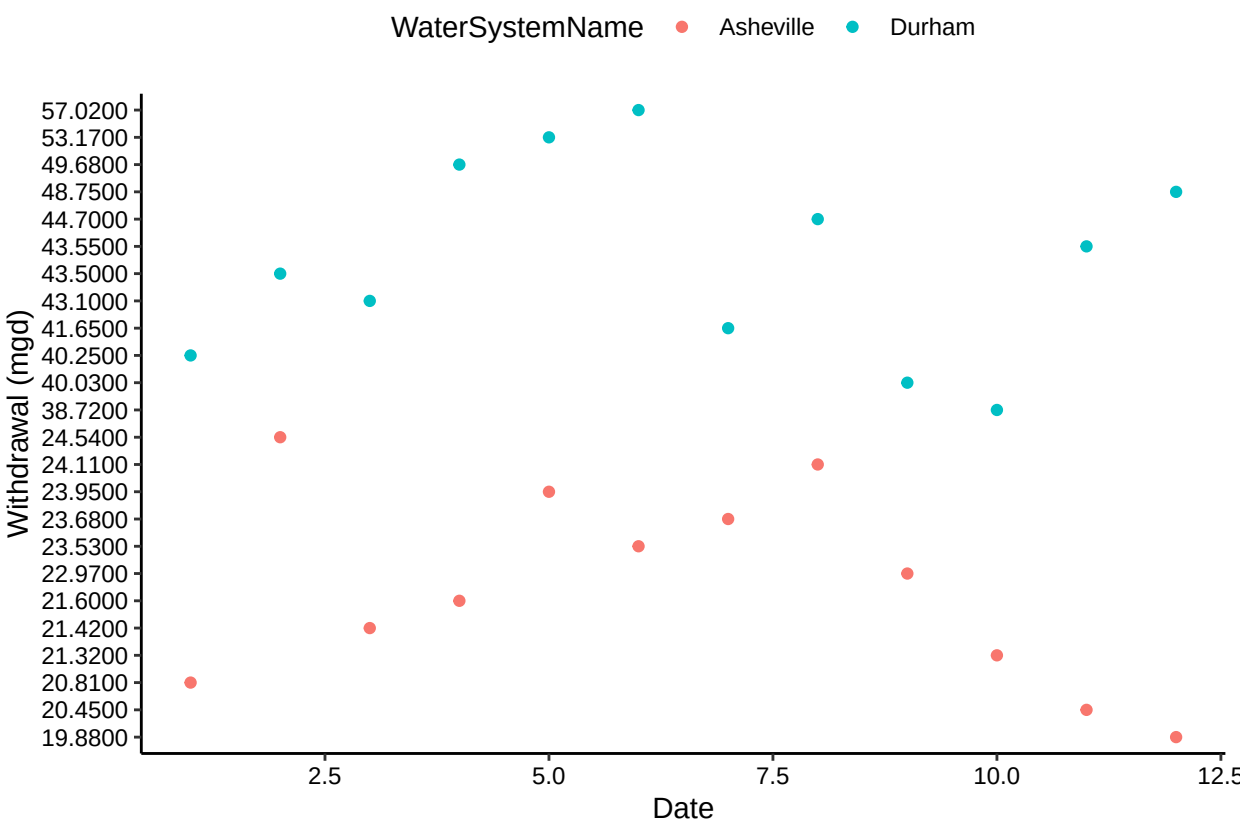


8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares Asheville's to Durham's water withdrawals.

```
#8
#Set the URL to be scraped, modifying PWSID and Year
website3 <- read_html('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2015')

#Scrape Water System Name
water.system.name3 <- html_nodes(website3,
                                 'div+ table tr:nth-child(1) td:nth-child(2)') %>%
  html_text()

#Scrape PWSID
PWSID3 <- html_nodes(website3, 'td tr:nth-child(1) td:nth-child(5)') %>%
  html_text()

#Scrape Ownership
ownership3 <- html_nodes(website3, 'div+ table tr:nth-child(2) td:nth-child(4)') %>%
  html_text()

#Scrape Maximum Day Use for each month
max.withdrawals.mgd3 <- html_nodes(website3, 'th~ td+ td') %>%
  html_text()
max.withdrawals.mgd
```

```
##  [1] "36.1000" "43.4200" "52.4900" "30.5000" "42.5900" "34.8800" "39.9100"
##  [8] "43.3200" "32.5300" "34.6600" "41.8000" "37.5300"
```

```
#Construct data frame for Asheville
ashevillewater.2015 <- data.frame(WaterSystemName = water.system.name3,
                          PWSID = PWSID3,
                          Ownership = ownership3,
                          MaxDayUse = max.withdrawals.mgd3,
                          Month = month(c(1,5,9,2,6,10,3,7,11,4,8,12)),
                          Year = rep(2015))

#Combine Asheville and Durham 2015 data frames
DurhamVAsheville <- rbind(ncwaterdf.2015, ashevillewater.2015)

#Plot comparative water withdrawals
DurhamVAsheville2015Plot <-
  ggplot(DurhamVAsheville, aes(x = Month,
                               y = MaxDayUse,
                               color = WaterSystemName)) +
  geom_point() +
  labs(y="Withdrawal (mgd)",
       x="Date")
  #geom_line()
print(DurhamVAsheville2015Plot)
```



9. 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 `bindrows()` to combine the dataframes into a single one.

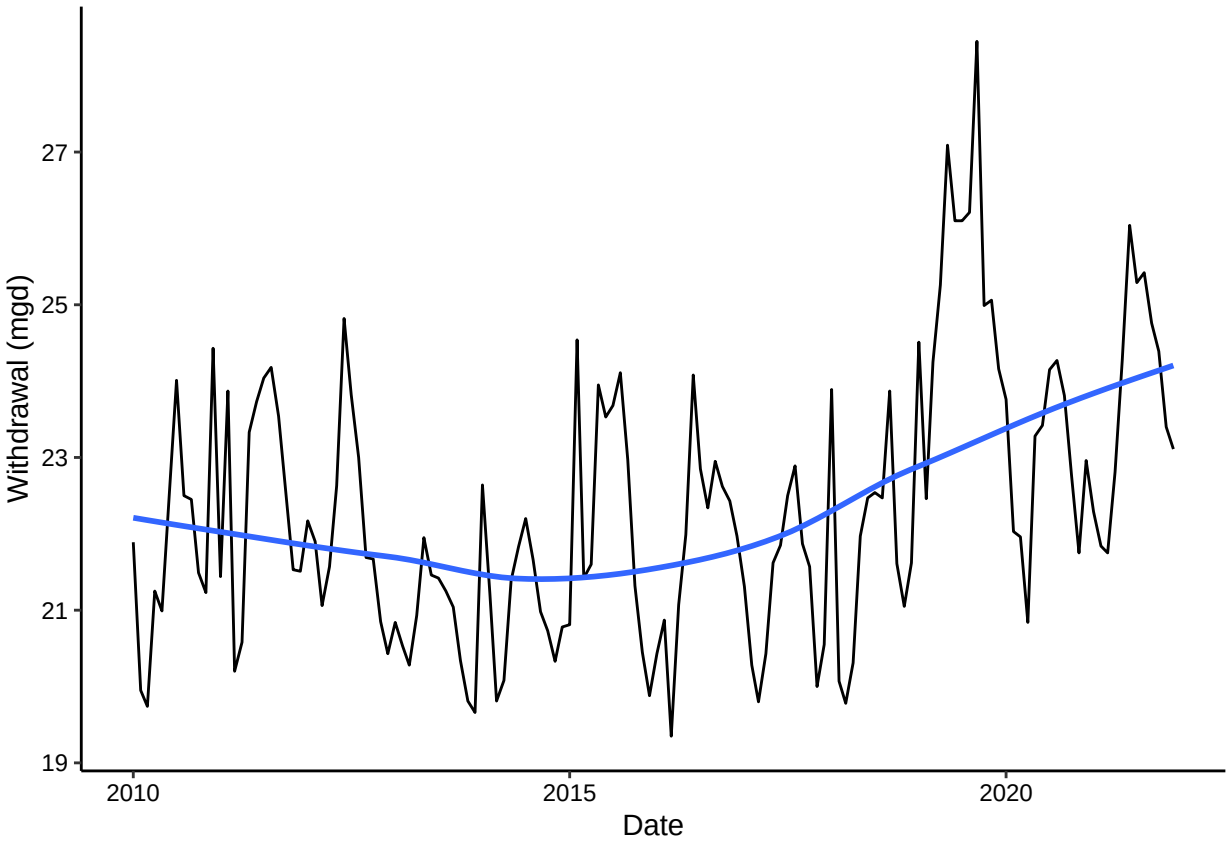
```
#9
#Construct scraping web address
the_base_url <- 'https://www.ncwater.org/WUDC/app/LWSP/'
the_facility <- '01-11-010'
the_year <- 2015
the_scrape_url <- paste0(the_base_url, 'report.php?pwsid=', the_facility,
                         '&year=', the_year)
print(the_scrape_url)
```

```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2015"
```

```
#Retrieve the website contents
the_website <- read_html(the_scrape_url)

#Set the element address variables
the_WaterSystemName_tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'
the_PWSID_tag <- 'td tr:nth-child(1) td:nth-child(5)'
the_Ownership_tag <- 'div+ table tr:nth-child(2) td:nth-child(4)'
the_MaxDayUse_tag <- 'th~ td+ td'

#Scrape the data items
the_WaterSystemName <- the_website %>% html_nodes(the_WaterSystemName_tag) %>%
  html_text()
the_PWSID <- the_website %>%   html_nodes(the_PWSID_tag) %>%  html_text()
the_Ownership <- the_website %>% html_nodes(the_Ownership_tag) %>% html_text()
the_MaxDayUse <- the_website %>% html_nodes(the_MaxDayUse_tag) %>% html_text()

#Construct a data frame from the scraped data
df_MaxDayUse <- data.frame("Month" = month(c(1,5,9,2,6,10,3,7,11,4,8,12)),
                           "Year" = rep(the_year,12),
                           "MaxDayUse" = as.numeric(the_MaxDayUse)) %>%
  mutate(Water_System_Name = !!the_WaterSystemName,
         PWSID = !!the_PWSID,
         Ownership = !!the_Ownership,
         Date = my(paste(Month,"-",Year)))

#Create our scraping function
scrape.it <- function(the_year, the_facility){

  #Retrieve the website contents
  the_website <- read_html(paste0(the_base_url, 'report.php?pwsid=',
                                  the_facility, '&year=', the_year))

  #Set the element address variables
  the_WaterSystemName_tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'
  the_PWSID_tag <- 'td tr:nth-child(1) td:nth-child(5)'
  the_Ownership_tag <- 'div+ table tr:nth-child(2) td:nth-child(4)'
```

```
  the_MaxDayUse_tag <- 'th~ td+ td'

  #Scrape the data items
  the_WaterSystemName <- the_website %>% html_nodes(the_WaterSystemName_tag) %>%
    html_text()
  the_PWSID <- the_website %>%   html_nodes(the_PWSID_tag) %>%  html_text()
  the_Ownership <- the_website %>% html_nodes(the_Ownership_tag) %>% html_text()
  the_MaxDayUse <- the_website %>% html_nodes(the_MaxDayUse_tag) %>% html_text()

  #Convert to a dataframe
  df_MaxDayUse <- data.frame("Month" = month(c(1,5,9,2,6,10,3,7,11,4,8,12)),
                             "Year" = rep(the_year,12),
                             "MaxDayUse" = as.numeric(the_MaxDayUse)) %>%
    mutate(Water_System_Name = !!the_WaterSystemName,
           PWSID = !!the_PWSID,
           Ownership = !!the_Ownership,
           Date = my(paste(Month,"-",Year)))

  Sys.sleep(1)

  return(df_MaxDayUse)
}

#Run the function
df_MaxDayUse <- scrape.it(2015,'01-11-010')
view(df_MaxDayUse)

#Set the inputs to scrape years 2010 to 2021 for the site "01-11-010"
the_years = rep(2010:2021)
my_facility = '01-11-010'

#Use purrr's map function
the_dfs <- map(the_years,scrape.it,the_facility=my_facility)

#Conflate the returned dataframes into a single dataframe
the_df <- bind_rows(the_dfs)

#Plot
ggplot(the_df,aes(x=Date,y=MaxDayUse)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(y="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: By looking at the plot, it appears that Asheville's maximum daily water withdrawals saw a slight overall decrease from 2010 to 2015, and an overall increase from 2015 to 2021. Over the period as a whole, Asheville's water usage has increased.