Assignment 09: Data Scraping

Meghan Seyler

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
# intialize packages
getwd()
```

```
## [1] "Z:/ENV872/Environmental_Data_Analytics_2022/Assignments"
```

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

```
## Warning: package 'rvest' was built under R version 4.1.3
```

```
library(dataRetrieval)
```

```
## Warning: package 'dataRetrieval' was built under R version 4.1.3
```

```
library(lubridate)

# set ggplot theme
my_theme09 <- theme_classic(base_size = 12) +
  theme(axis.text = element_text(color = "black"),
  legend.position = "right")
theme_set(my_theme09)
```

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
# read in url
theURL <- read_html('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2020')
theURL
```

```
## {html_document}
## <html xmlns="http://www.w3.org/1999/xhtml" lang="en" xml:lang="en">
## [1] <head>\n<title>DWR :: Local Water Supply Planning</title>\n<meta http-equ ...
## [2] <body id="plan">\r\n<!--<div id="division-header">\r\n<a name="top" href= ...
```

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

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

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

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

# scrape max withdrawal values
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.)

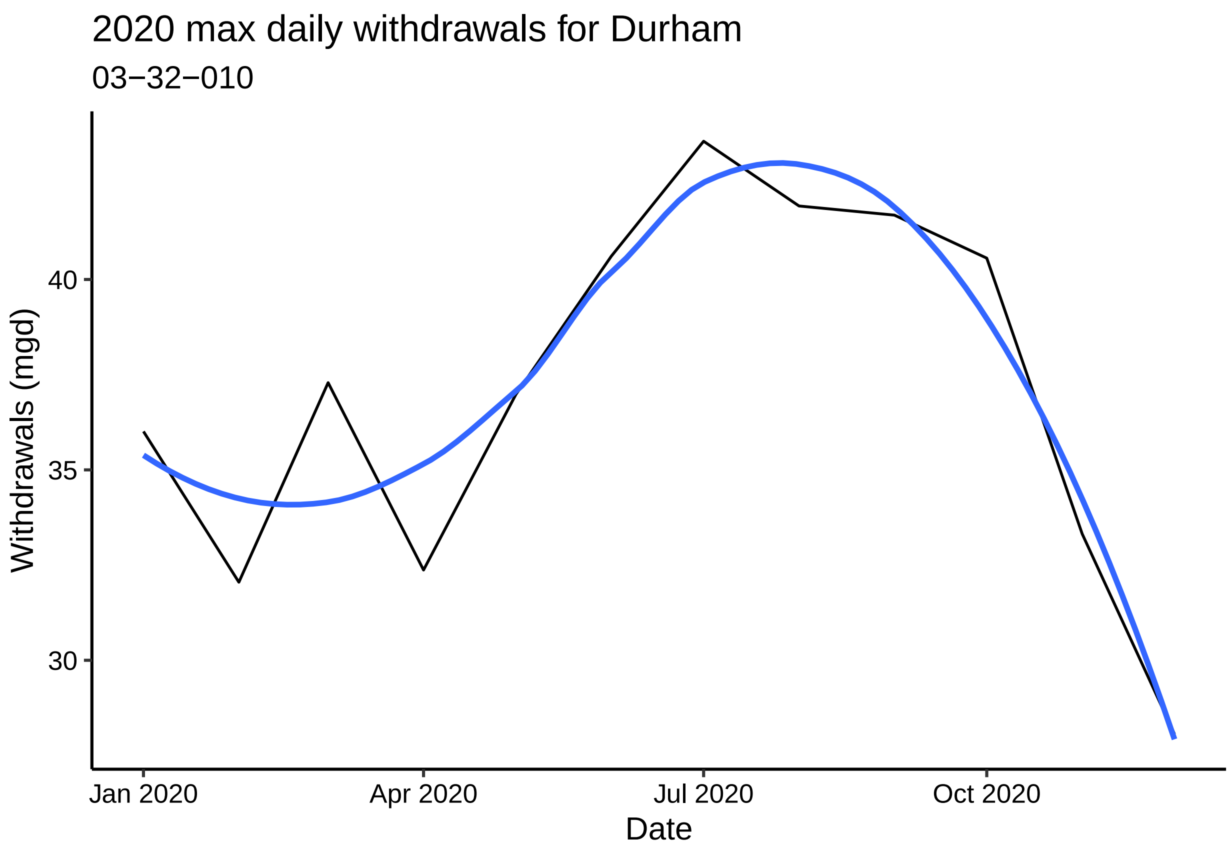
   TIP: Use `rep()` to repeat a value when creating a dataframe.

   NOTE: It's likely you won't be able to scrape the monthly 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

```r
#4
# create dataframe
df_LWSP <- data.frame("Month" = c("Jan", "May", "Sep", "Feb", "Jun", "Oct", "Mar", "Jul", "Nov", "Apr",
                      "Year" = rep(2020, 12),
                      "Max_day_use" = as.numeric(max.withdrawals.mgd)) %>%
# populate values
    mutate(System_name = !!water.system.name,
         PWSID = !!pswid,
         Ownership = !!ownership,
         Date=my(paste(Month,"-", Year)))
#5
# plot max withdrawals for Durham in 2020
max_withdrawl_plot<-ggplot(df_LWSP,aes(x=Date, y=Max_day_use)) +
  geom_line() +
geom_smooth(method="loess", se=FALSE) +
  labs(title=paste("2020 max daily withdrawals for Durham"),
       subtitle = pswid,
       y="Withdrawals (mgd)",
       x="Date")
max_withdrawl_plot
```

```
## `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.
# create the scraping function
scrape.it <- function(the_year, the_code){
  the_url <- paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=', the_code, '&year=', the_
  print(the_url)

# gather website content
  webpage <- read_html(the_url)
# set the element tags
  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)'
  withdrawals.tag <- 'th~ td+ td'
# scrape the data items
  name.scraped <- webpage %>% html_nodes(system.name.tag) %>% html_text()
  pwsid.scraped <- webpage %>% html_nodes(pwsid.tag) %>% html_text()
  ownership.scraped <- webpage %>% html_nodes(ownership.tag) %>% html_text()
  withdrawal.scraped <- webpage %>% html_nodes(withdrawals.tag) %>% html_text()

# convert to a dataframe
  the_df<-data.frame("Month" = c("Jan", "May", "Sep", "Feb", "Jun", "Oct", "Mar", "Jul", "Nov", "Apr",
                     "Year" = rep(the_year, 12),
                      "Max_withdrawals_mgd" = as.numeric(withdrawal.scraped)) %>%
    mutate('WaterSystem' = as.character(name.scraped),
                  "PWSID" = as.character(pwsid.scraped),
                  "Ownerhip" = as.character(ownership.scraped),
                  "Max_withdrawals_mgd" = as.numeric(withdrawal.scraped),
            "Date" = my(paste(Month,"-",Year)))
# return the dataframe
  return(the_df)
}
```

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

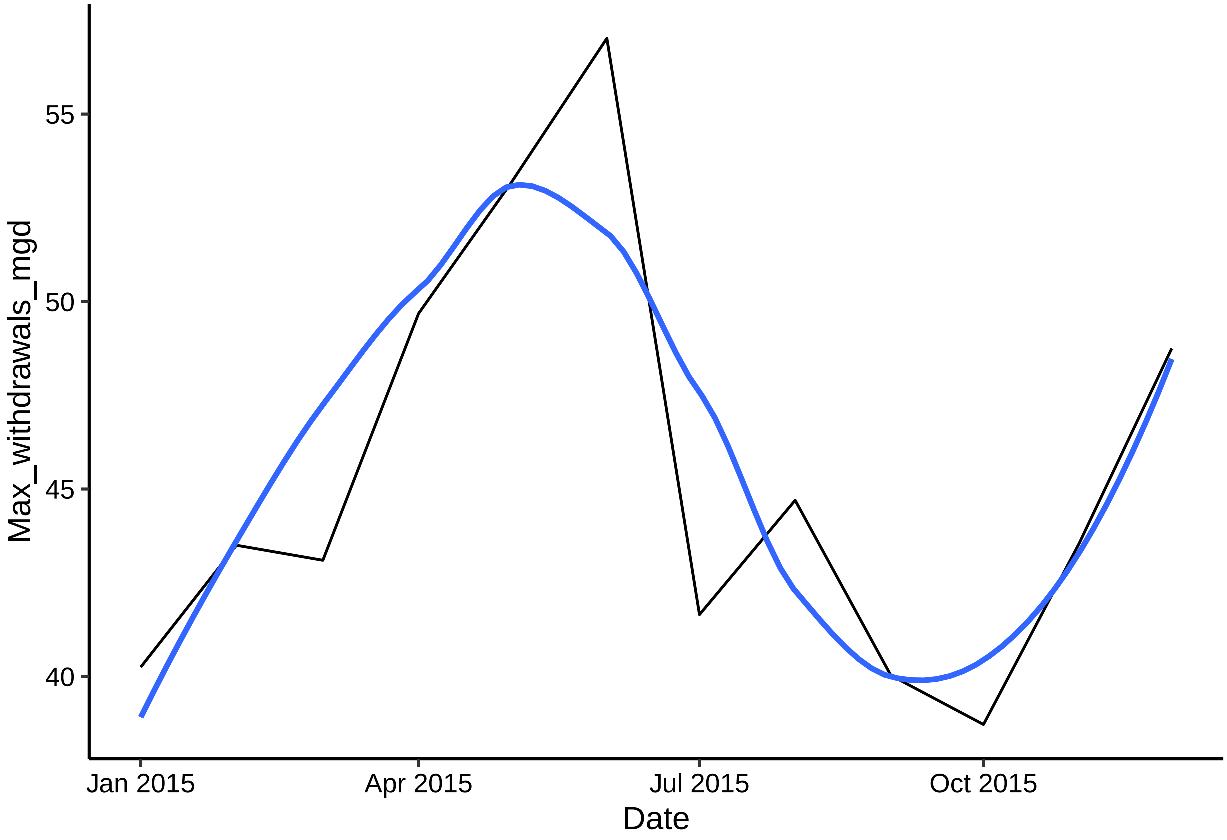
```
#7
# apply function
the_df <- scrape.it(2015,'03-32-010')
```

```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2015"
```

```
view(the_df)

# plot data
ggplot(the_df,aes(x=Date, y=Max_withdrawals_mgd)) +
  geom_line()+
  geom_smooth(method="loess",se=FALSE)
```

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



```
  labs(title = "2015 Durham Max Daily Withdrawals by Month",
       x = "Date",
       y = "Max Withdrawals (MGD)")
```

```
## $x
## [1] "Date"
##
## $y
## [1] "Max Withdrawals (MGD)"
##
## $title
## [1] "2015 Durham Max Daily Withdrawals by Month"
##
## attr(,"class")
## [1] "labels"
```

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
# apply function - Asheville 2015
the_df2 <- scrape.it(2015,'01-11-010')
```

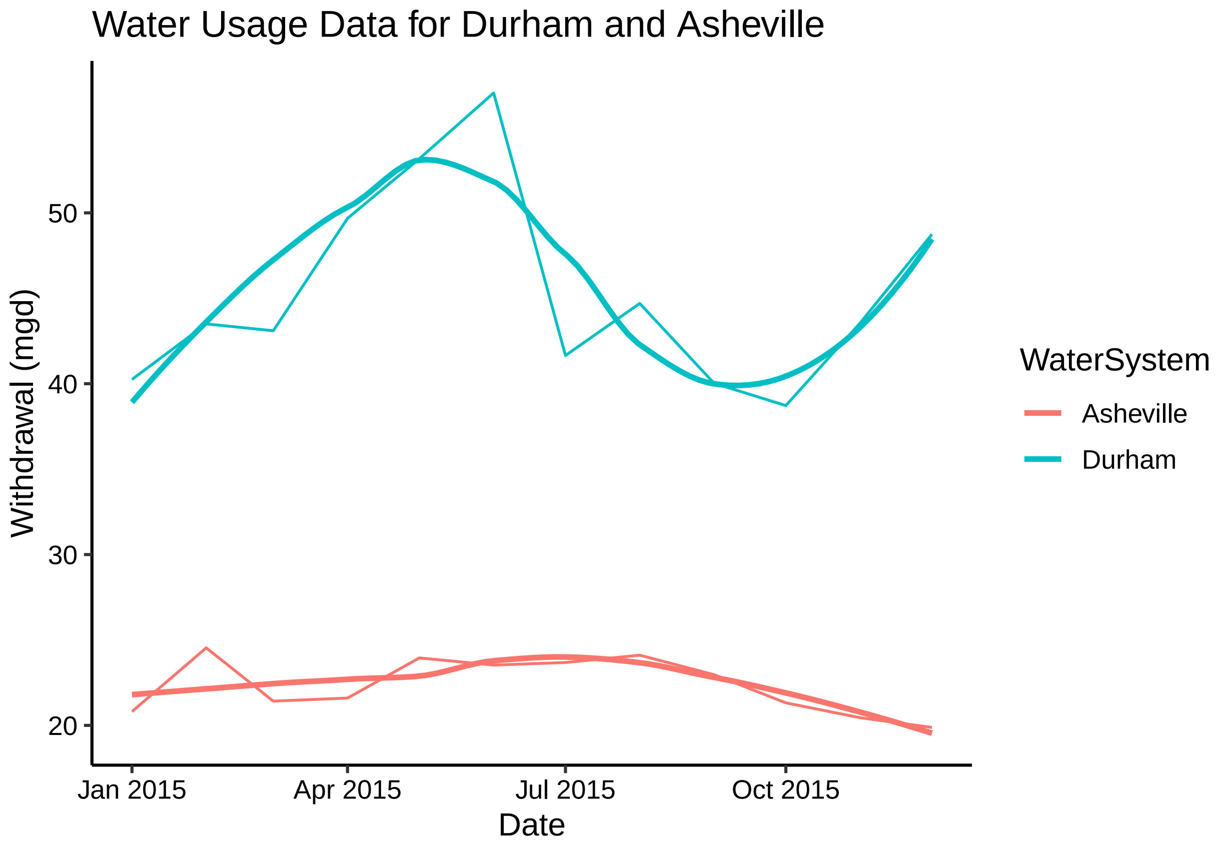
```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2015"
```

```
view(the_df2)

# combined datasets
df_combined <- bind_rows(the_df, the_df2)
view(df_combined)

# plot Durham vs. Asheville
ggplot(df_combined,aes(x=Date,y=Max_withdrawals_mgd, color = WaterSystem)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = "Water Usage Data for Durham and Asheville",
       y="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.

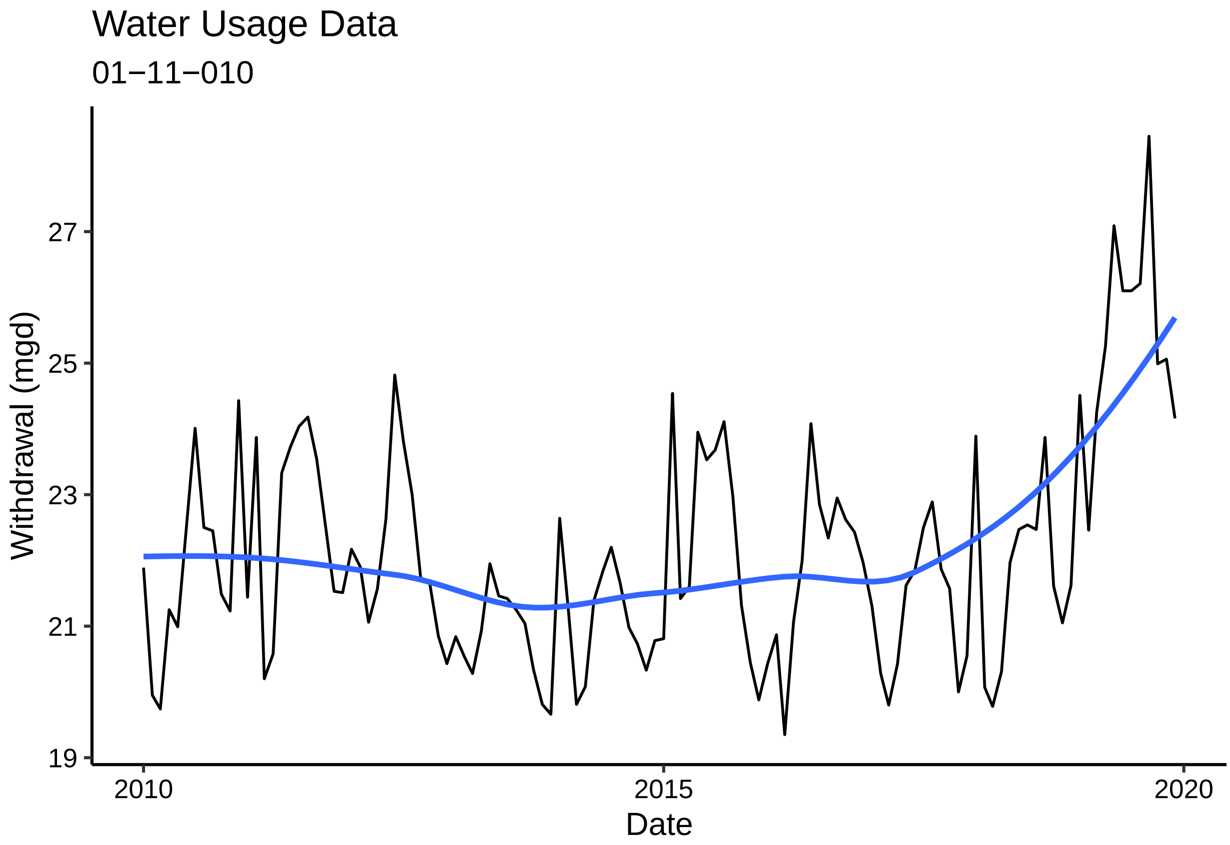
```
#9
the_year = seq(2010,2019)
the_code = '01-11-010'

asheville_data <- the_year %>%
  map(scrape.it, the_code) %>%
  bind_rows()
```

```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2010"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2011"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2012"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2013"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2014"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2015"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2016"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2017"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2018"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2019"
```

```
ggplot(asheville_data,aes(x=Date,y=Max_withdrawals_mgd)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = "Water Usage Data",
       subtitle = the_code,
       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?

Yes, there is a positive, upward sloping trend in water use in Asheville for the 2010-2019 period.