Assignment 10: Data Scraping

Laura Exar

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
#Loaded packages
library(tidyverse);library(rvest); library(ggplot2); library(lubridate)

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

```r
#2
#Indicated the website as the URL to be scraped
the_url <- 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 “27.6400”.

```r
#3
#Collected data and assigned them to variables
water.system.name <- the_url %>% html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>% html_text

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

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

max.withdrawals.mgd <- the_url %>% html_nodes('th~ td+ td') %>% html_text()
```

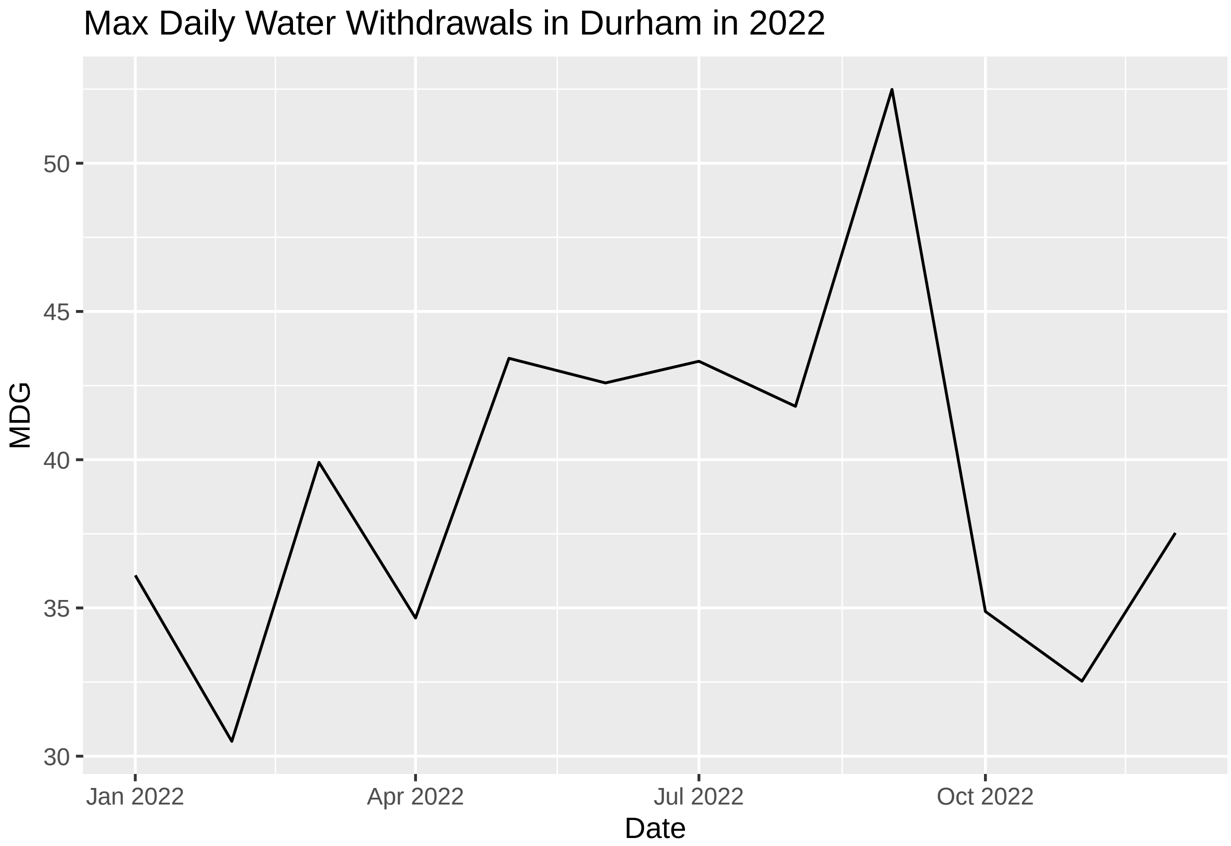
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

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

   NOTE: It’s likely you won’t be able to scrape the monthly 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 average daily withdrawals across the months for 2022

```
#4
#Created a dataframe from the variables
Month <- c('Jan', 'May', 'Sep', 'Feb', 'Jun', 'Oct', 'Mar', 'Jul', 'Nov', 'Apr', 'Aug', 'Dec')

withdrawal_df <- data.frame(
  "WaterSystemName" = water.system.name,
  "PWSID" = PWSID,
  "Ownership" = ownership,
  "Month" = Month,
  "Date" = my(paste0(Month,"-","2022")),
  "MaxDayUse" = as.numeric(max.withdrawals.mgd))

withdrawal_df <- arrange(withdrawal_df, Date)

#5
#Created a line plot
Durham2022 <- ggplot(withdrawal_df, aes(x=Date,y=MaxDayUse)) +
  geom_line() +
  labs(x = "Date", y = "MDG", title = "Max Daily Water Withdrawals in Durham in 2022")
Durham2022
```



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**.

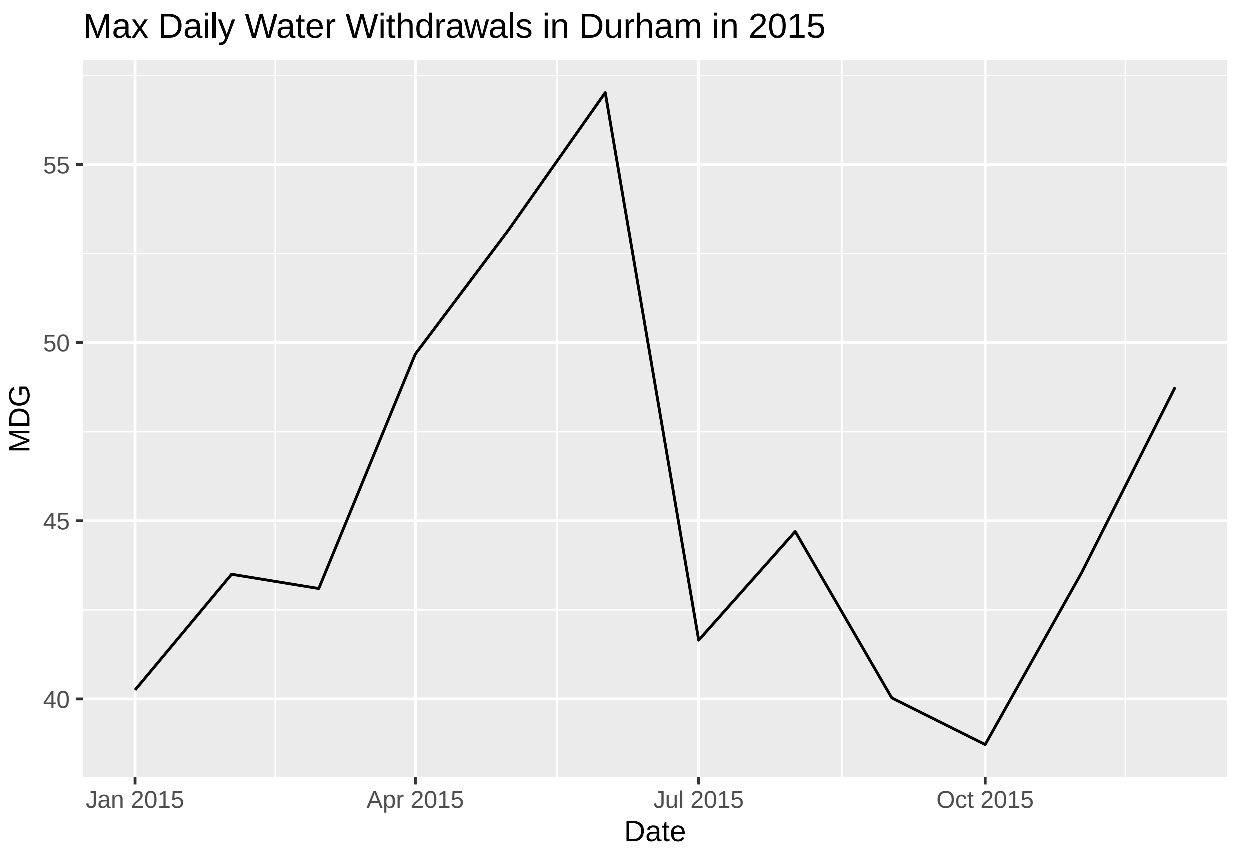
```r
#6.
#Created a scraping function
scrape.it <- function(PWSID, the_year){
  #Get the proper url
the_url <- paste0(
  'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',
  PWSID,
  '&year=',
  the_year)
print(the_url)

#Fetch the website
the_website <- read_html(the_url)

#Collected data and assigned them to variables
water.system.name <- the_website %>% html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>% html_t
PWSID <- 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()

#Created a dataframe from the variables
Month <- c('Jan', 'May', 'Sep', 'Feb', 'Jun', 'Oct', 'Mar', 'Jul', 'Nov', 'Apr', 'Aug', 'Dec')

withdrawal_df <- data.frame(
  "WaterSystemName" = water.system.name,
  "PWSID" = PWSID,
  "Ownership" = ownership,
  "Month" = Month,
  "Date" = my(paste0(Month,"-",the_year)),
  "MaxDayUse" = as.numeric(max.withdrawals.mgd))

withdrawal_df <- arrange(withdrawal_df, Date)

#Return the dataframe
return(withdrawal_df)
}
```

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

```r
#7
#Extracted max daily withdrawals for Durham for 2015
dfDurham <- scrape.it("03-32-010","2015")
```

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

```r
#Plotted max daily withdrawals for Durham for 2015
DurhamPlot <- ggplot(dfDurham, aes(x=Date,y=MaxDayUse)) +
  geom_line() +
  labs(x = "Date", y = "MDG", title = "Max Daily Water Withdrawals in Durham in 2015")
DurhamPlot
```

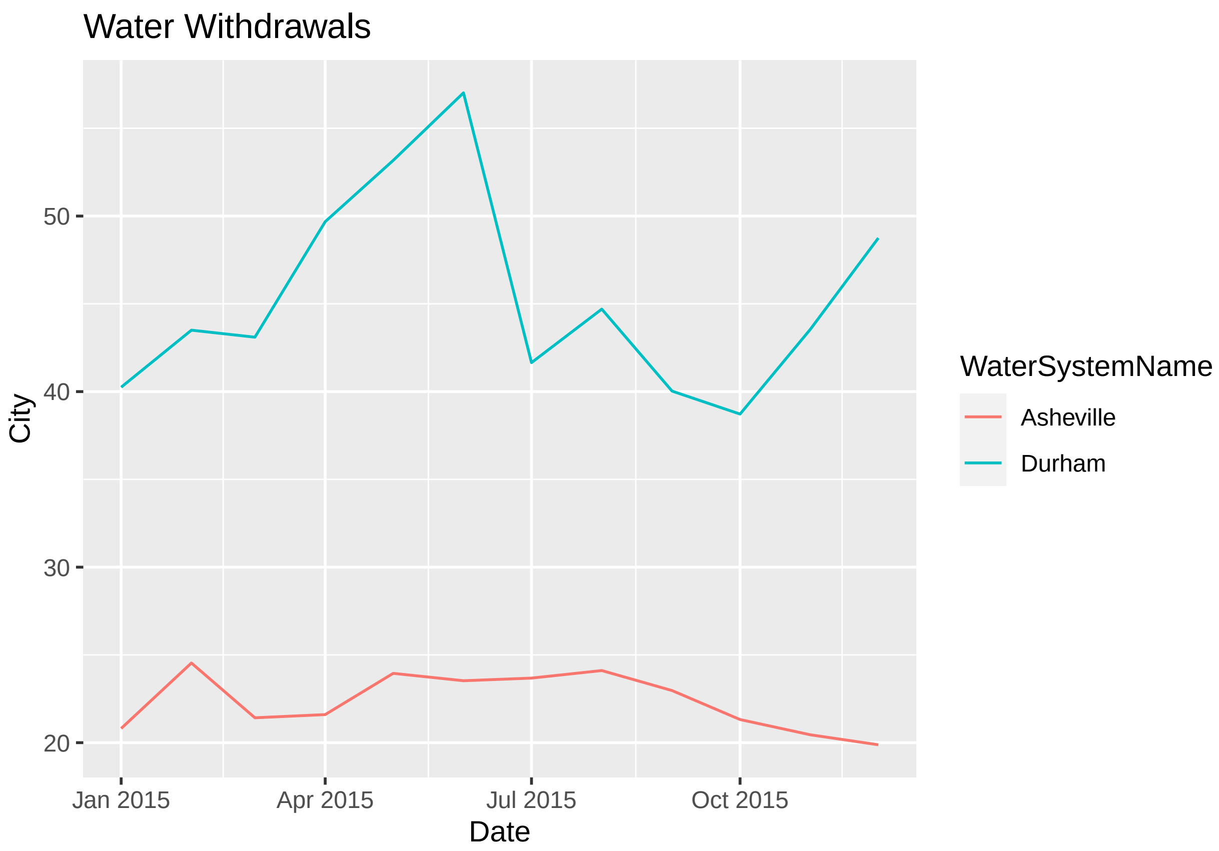


8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares Asheville's to Durham's water withdrawals.

```
#8
#Extracted data for Asheville in 2015
dfAsheville <- scrape.it("01-11-010","2015")
```

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

```
#Combined Asheville data with Durham
df2015 <- rbind(dfDurham, dfAsheville)

#Created a plot that compares Asheville to Durham
Plot2015 <- df2015 %>%
  ggplot(aes(x=Date, y=MaxDayUse, color= WaterSystemName)) +
  geom_line() +
  labs(title="Water Withdrawals",
       x="Date",
       y="City")
Plot2015
```



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').

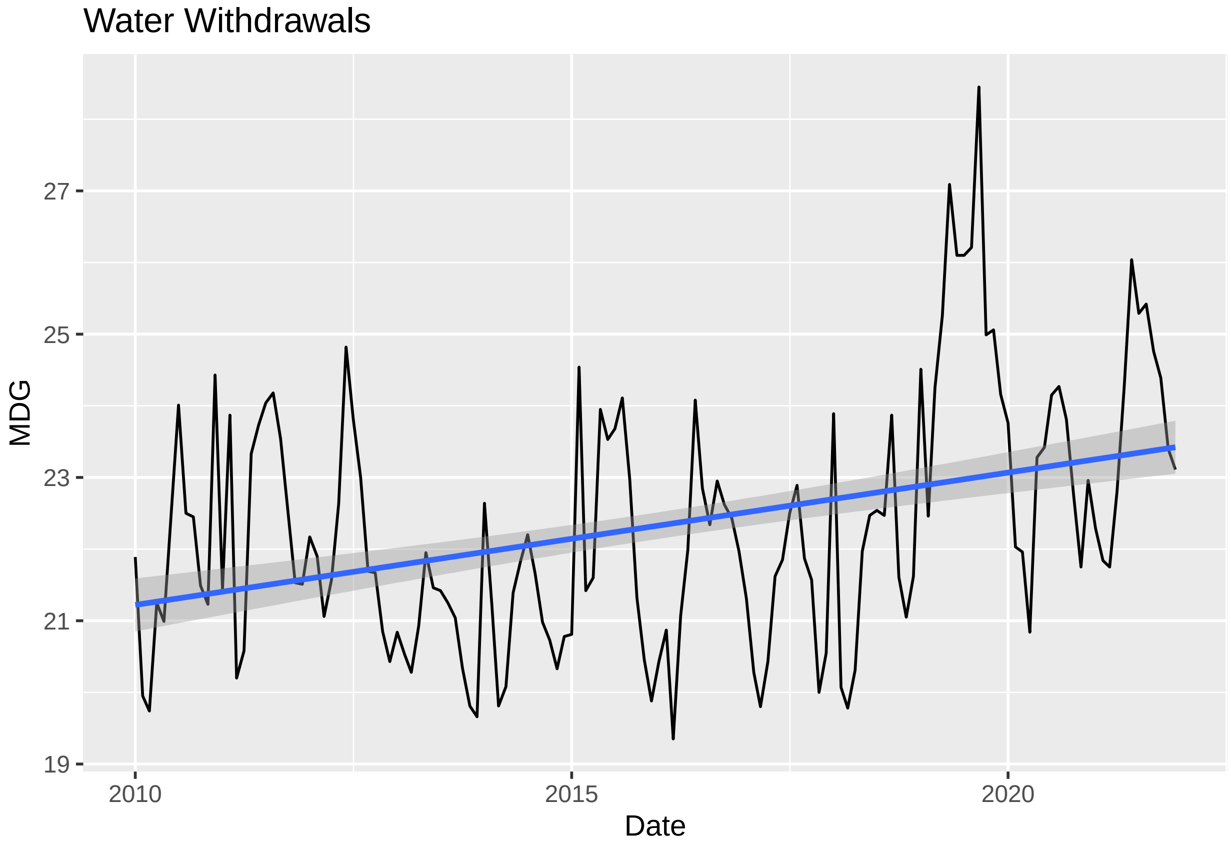
   TIP: See Section 3.2 in the "09_Data_Scraping.Rmd" where we apply "map2()" to iteratively run a function over two inputs. Pipe the output of the map2() function to `bindrows()` to combine the dataframes into a single one.

```
#9
#Created a dataframe of Asheville's max daily withdrawals for 2010 to 2021
the_years = c(2010,2011,2012,2013,2014,2015,2016,2017,2018,2019,2020,2021)
facility = rep("01-11-010",length(the_years))

dfs_Asheville <- map2(facility, the_years, scrape.it)
```

```
## [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"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2020"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2021"
```

```
dfs_Asheville <- dfs_Asheville %>%
  bind_rows(dfs_Asheville)

#Plotted Asheville's max daily withdrawals for 2010 to 2021
AshevillePlot <- dfs_Asheville %>%
  ggplot(aes(x=Date, y=MaxDayUse)) +
  geom_line() +
  geom_smooth(method = lm) +
  labs(title="Water Withdrawals",
       x="Date",
       y="MDG")
AshevillePlot
```

```
## ‘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? Looking at this plot, Asheville’s water withdrawals has a trend of water usage increasing over time.