Assignment 10: Data Scraping

Karina Leung

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
library(tidyverse);library(lubridate);library(viridis);library(here)
library(rvest)
library(dataRetrieval)
library(tidycensus)

mytheme <- theme_classic(base_size = 12) +
  theme(axis.text = element_text(color = "black", size = 12),
        plot.title.position = "plot",
        plot.title = element_text(size = 16, face = "bold", hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        axis.ticks = element_line(color = "black", linewidth = 0.5),
        legend.background = element_rect(color='grey', fill = 'white'),
        legend.title = element_text(color='black', hjust = 0.5),
        legend.position = "right")
theme_set(mytheme)
```

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.)

```
#2
NC.Water.Municipality.Webpage <-
↪  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".

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

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

```
PWSID <- NC.Water.Municipality.Webpage %>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
  html_text()
PWSID
```

```
## [1] "03-32-010"
```

```
ownership <- NC.Water.Municipality.Webpage %>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%
  html_text()
ownership
```

```
## [1] "Municipality"
```

```
max.withdrawals.mgd <- NC.Water.Municipality.Webpage %>%
  html_nodes("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"
```

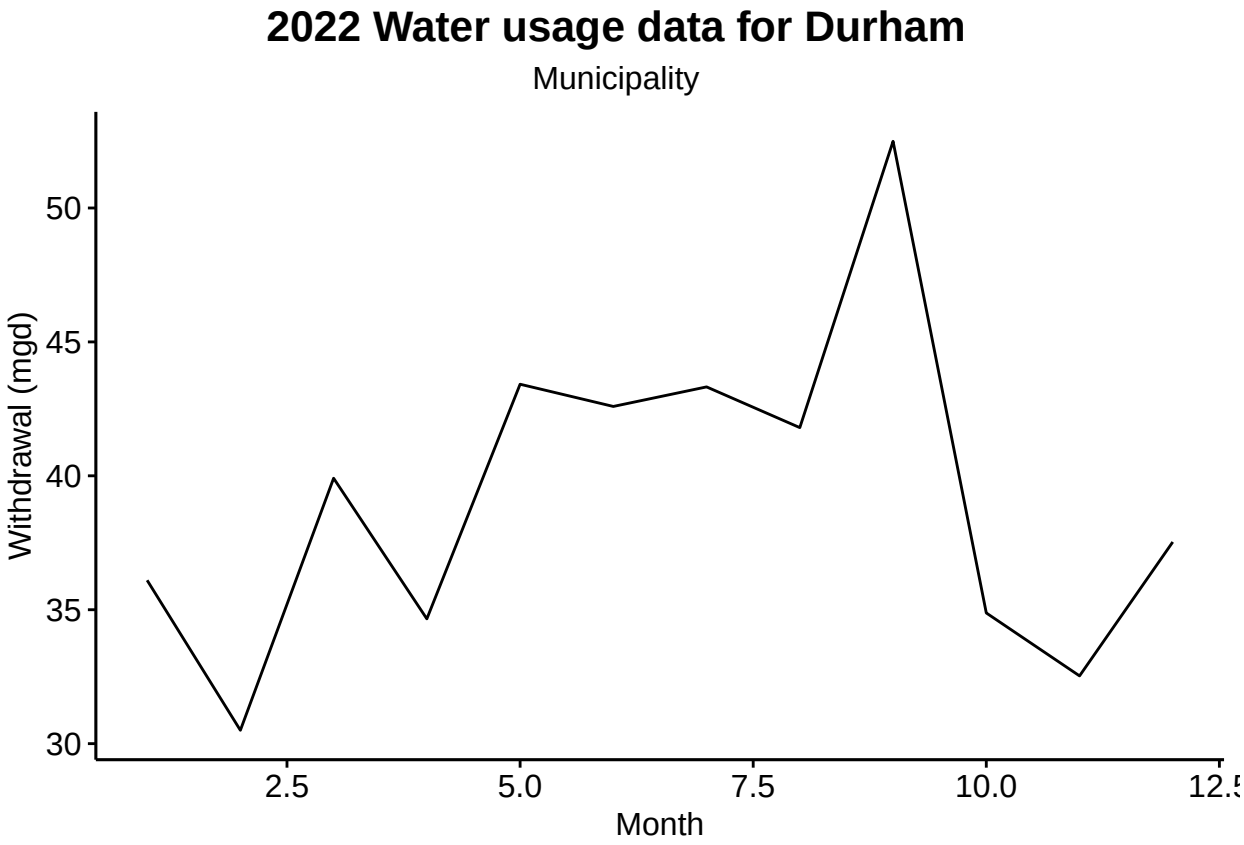
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   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
water.withdrawals.df <- data.frame("Month" = c(1, 5, 9, 2, 6, 10, 3, 7, 11, 4, 8, 12),
                                   "Year" = rep(2022),
                                   "Max_Withdrawals_mgd" = as.numeric(max.withdrawals.mgd))


#5
ggplot(water.withdrawals.df, aes(x=Month,y=Max_Withdrawals_mgd)) +
  geom_line() +
 # geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2022 Water usage data for",water.system.name),
       subtitle = ownership,
       y="Withdrawal (mgd)",
       x="Month")
```



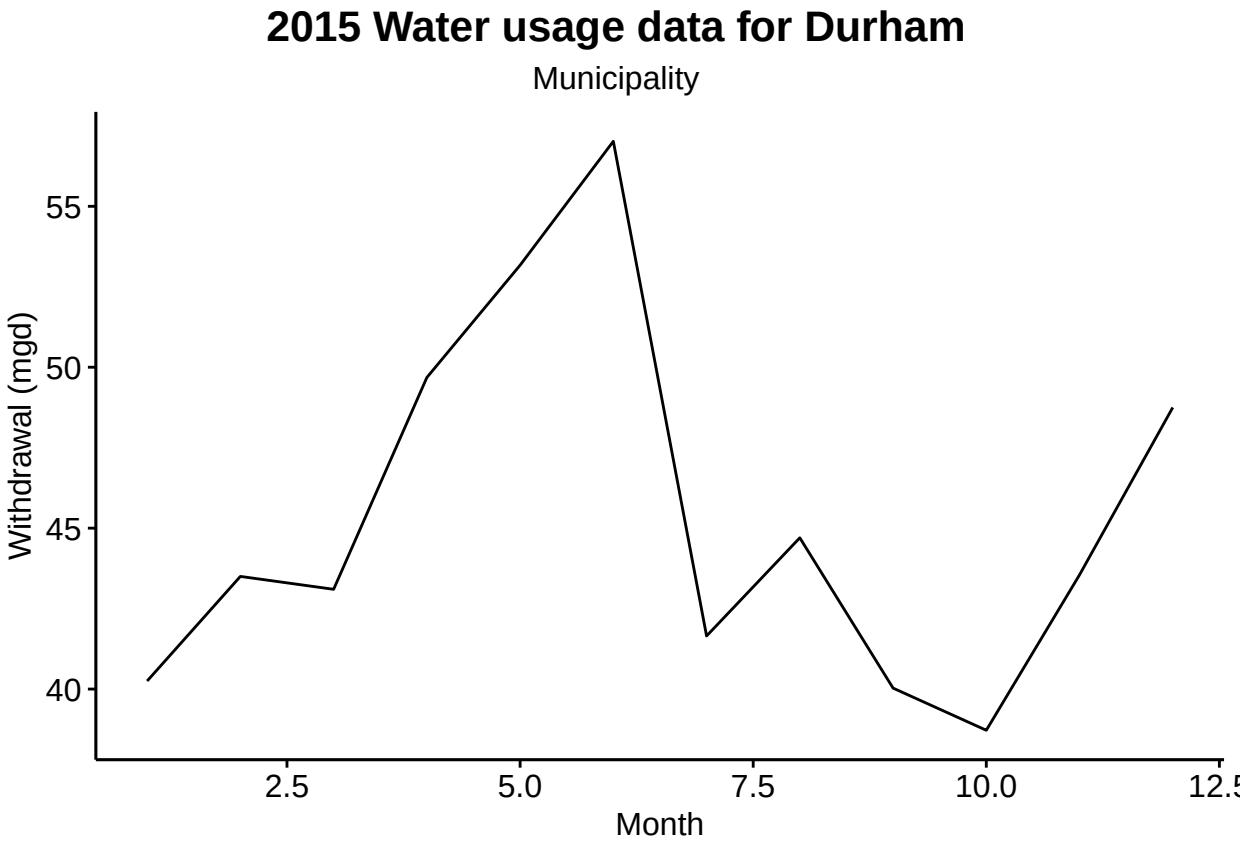
6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

```
#6.
scrape.it <- function(the_year, pwsid){
   #Fetch the website
  the_website <-
↪  read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=', pwsid,
↪  '&year=', the_year))

  #Scrape the data
  water.system.name <- the_website %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()

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()

  #Convert to dataframe
  functionfor.water.withdrawals.df <- data.frame("Month" = c(1, 5, 9, 2, 6, 10, 3, 7, 11,
↪  4, 8, 12),
                                "Year" = rep(the_year,12),
                                "Max_Withdrawals_mgd" = as.numeric(max.withdrawals.mgd)) %>%
     mutate(Water.System.Name = !!water.system.name,
            PWSID= !!PWSID,
            Ownership = !!ownership,
            Date = my(paste(Month,"-",Year)))

  #Return the dataframe
  return(functionfor.water.withdrawals.df)
}
```

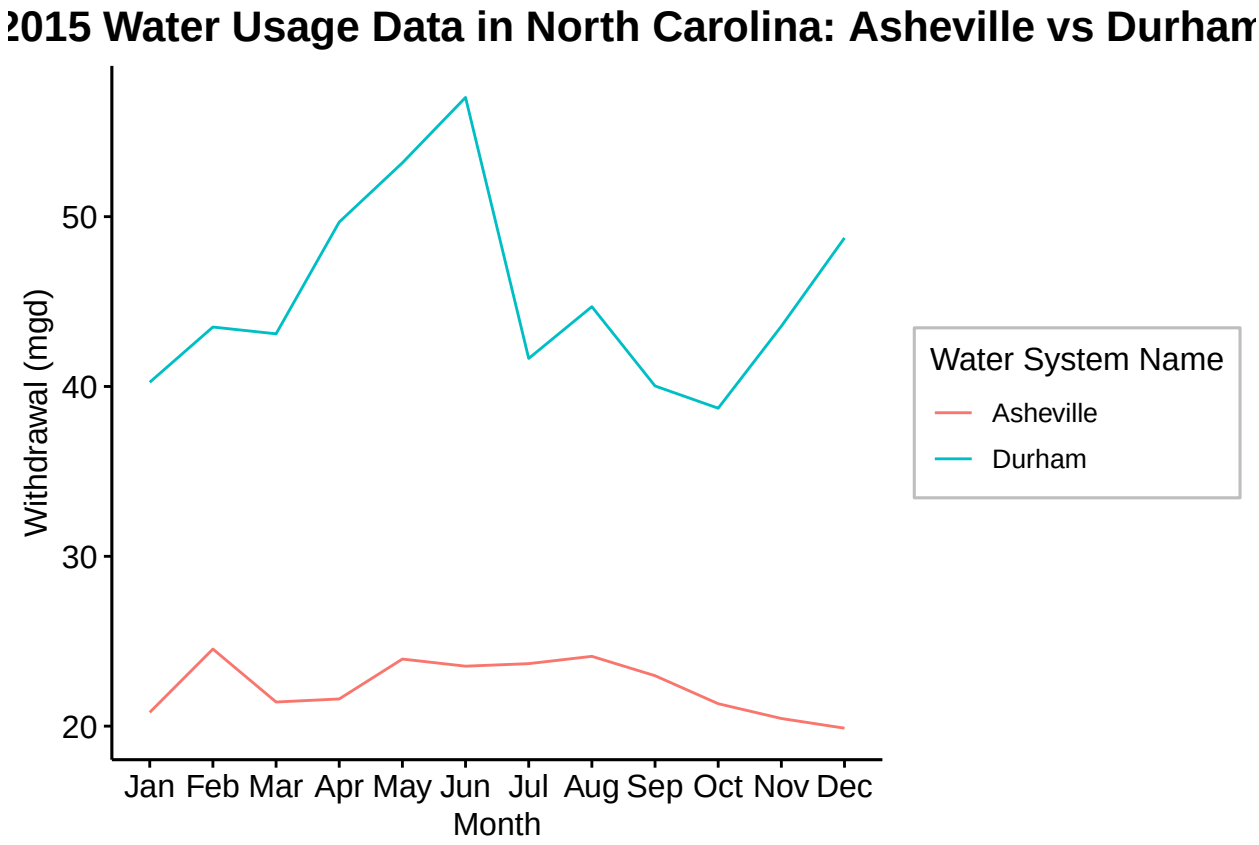
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID='03-32-010') for each month in 2015

```
#7
Durham.waterwithdrawal.df <- scrape.it(2015,'03-32-010')
view(Durham.waterwithdrawal.df)

ggplot(Durham.waterwithdrawal.df, aes(x=Month,y=Max_Withdrawals_mgd)) +
  geom_line() +
 # geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2015 Water usage data for",water.system.name),
       subtitle = ownership,
       y="Withdrawal (mgd)",
       x="Month")
```



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
Asheville.waterwithdrawal.df <- scrape.it(2015,'01-11-010')
view(Asheville.waterwithdrawal.df)

DurhamAsheville.waterwithdrawal.df <- rbind(Durham.waterwithdrawal.df,
↪  Asheville.waterwithdrawal.df)

ggplot(DurhamAsheville.waterwithdrawal.df, aes(x=factor(Month, levels = 1:12, labels =
↪  month.abb), y=Max_Withdrawals_mgd, color=Water.System.Name, group=Water.System.Name))
↪  +
  geom_line() +
  labs(title = "2015 Water Usage Data in North Carolina: Asheville vs Durham",
       y="Withdrawal (mgd)",
       x="Month",
       color="Water System Name")
```

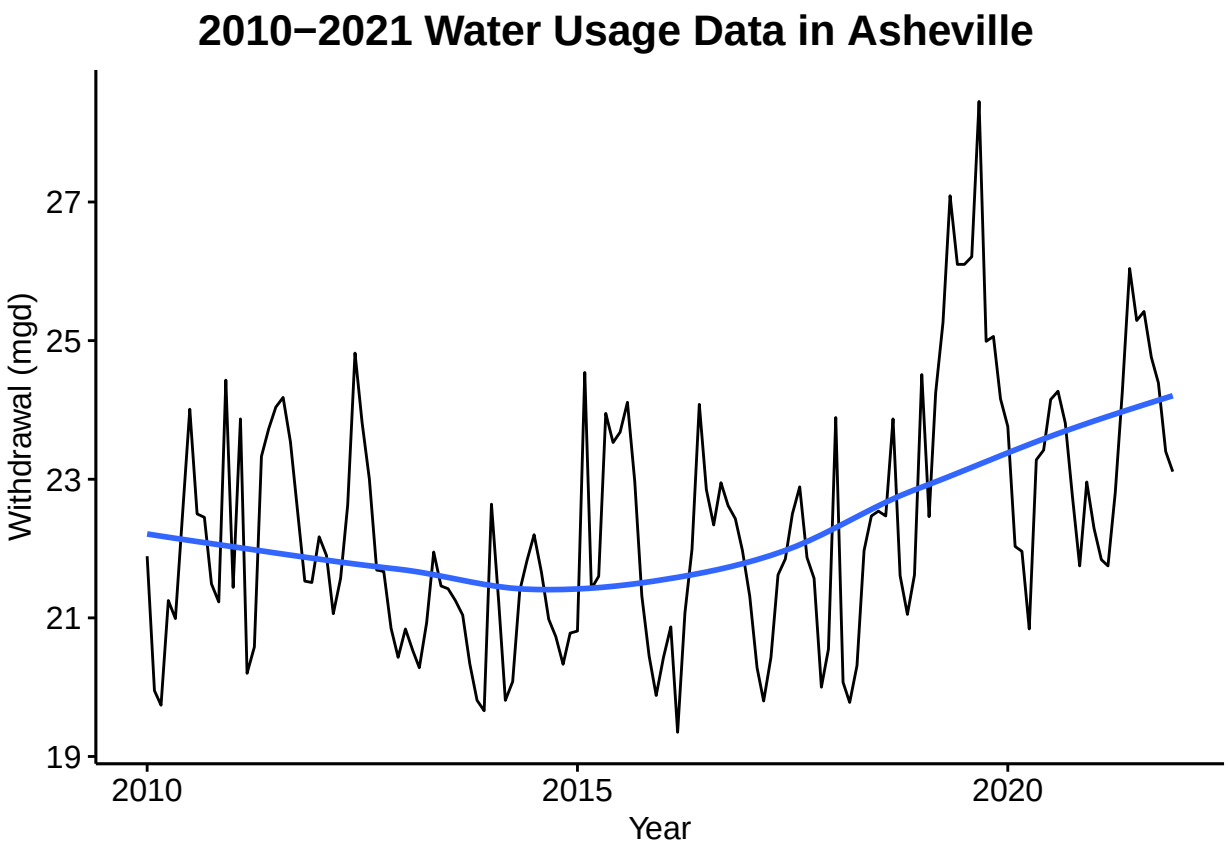


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2021.Add a smoothed line to the plot (method = 'loess').

   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
#Pick out Asheville years
Asheville.years <-  c(2010:2021)

#Create a list of the PWSID we want, the same length as the vector above
Asheville_PWSID <- rep.int('01-11-010', length(Asheville.years))

#"Map" the "scrape.it" function to retrieve data for all these
Asheville.Decade.Withdrawals <- map2(Asheville.years, Asheville_PWSID, scrape.it)

#Combine the returned list of dataframes into a single one
Asheville.Decade.Withdrawals <- bind_rows(Asheville.Decade.Withdrawals)


#Plot
ggplot(Asheville.Decade.Withdrawals, aes(x=Date, y=Max_Withdrawals_mgd)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
```

```
labs(title = "2010-2021 Water Usage Data in Asheville",
     y="Withdrawal (mgd)",
     x="Year")
```



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 trend in water usage over time. The trend line shows a decrease in water usage from 2010-2015, and then an increase in water usage to 2021 above 2010 levels.