Assignment 10: Data Scraping

Srishti Mutha

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(here); here()
```

```
## [1] "/Users/cherry/Desktop/EDA/EDA-Spring2023"
```

```
library(dplyr)

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

getwd()
```

```
## [1] "/Users/cherry/Desktop/EDA/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.)

```
#2
webpage <- read_html('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2022')
webpage
```

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

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

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

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

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

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

```
month <- webpage%>%
  html_nodes(".fancy-table:nth-child(31) tr+ tr th") %>%
  html_text()
month
```

```
##  [1] "Jan" "May" "Sep" "Feb" "Jun" "Oct" "Mar" "Jul" "Nov" "Apr" "Aug" "Dec"
```

```
max.withdrawals.mgd <- 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.

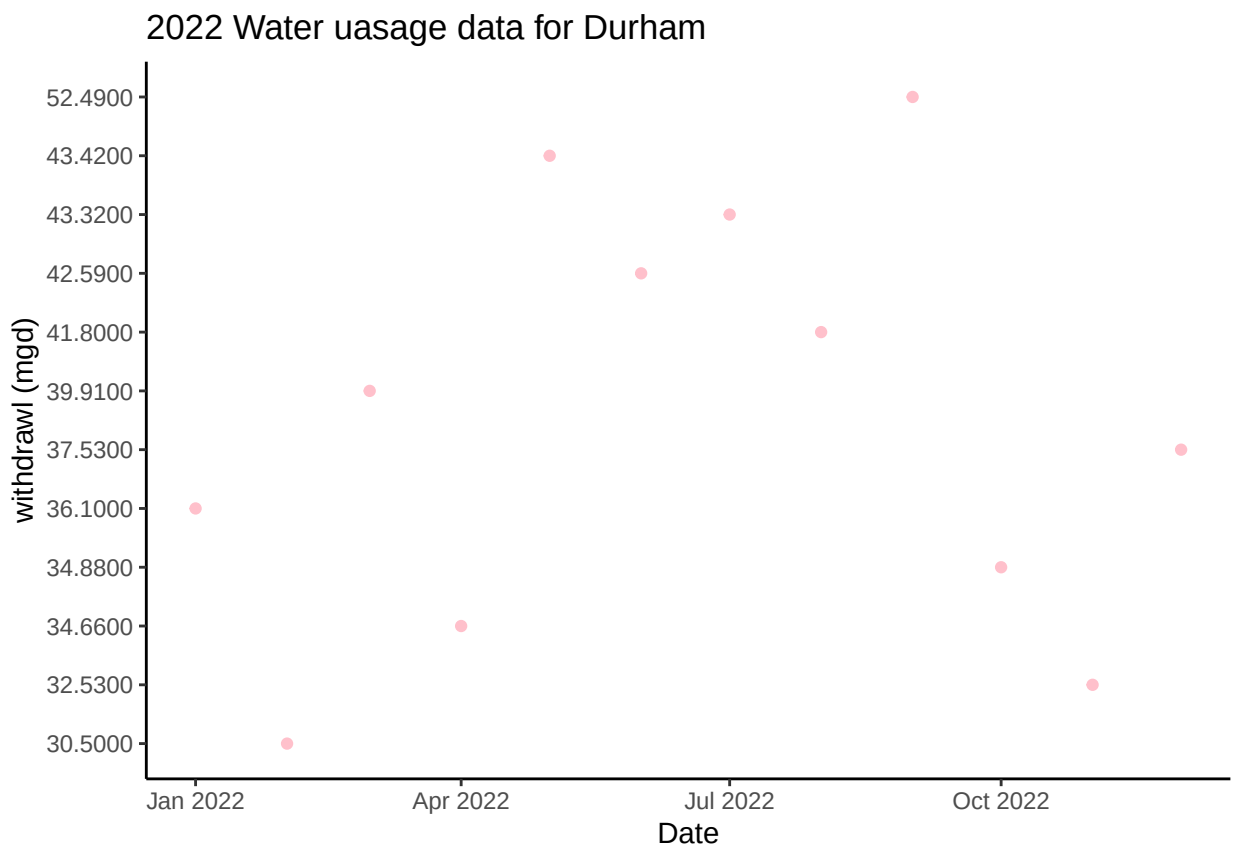
   NOTE: It's likely you won't be able to scrape the monthly widthrawal data in chronological order. You can overcome this by creating a month column manually assigning values in the order the data are scraped: "Jan", "May", "Sept", "Feb", etc... Or, you could scrape month values from the web page...

5. Create a line plot of the average daily withdrawals across the months for 2022

```
#4
withdrawl <-data.frame("year" = rep(2022, times = 12),
                       "Max_withdrawals.mgd" = as.numeric(max.withdrawals.mgd))

withdrawl <- withdrawl%>%
  mutate(water_system_name = !!water.system.name,
         PWSID = !!PWSID,
         ownership = !!ownership,
         Month = !!month,
         Date = my(paste(month, "-", year)))

#5
withdrawl_plot <- ggplot(withdrawl, aes(x=Date, y = max.withdrawals.mgd))+
  geom_line()+
```

```
  geom_point(color= "pink") +
  labs(title = paste("2022 Water uasage data for", water.system.name),
       y="withdrawl (mgd)",
       x= "Date") +
  theme_classic()
withdrawl_plot
```

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



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**.

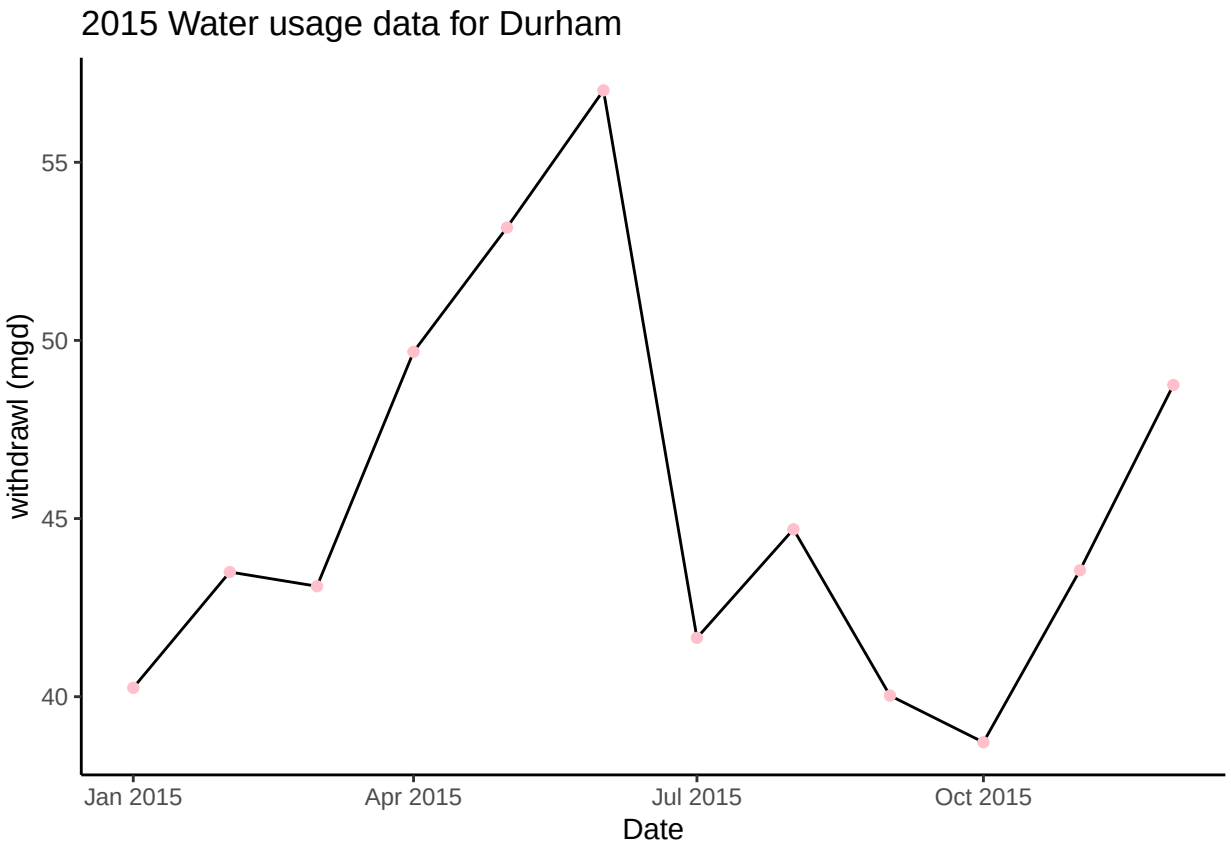
```
#6.


scrape.it <- function(the_year, the_pwsid){

  #Retrieve the website contents
  the_website <- read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?', 'pwsid=',the_pws

  the_water.system.name_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_max_mgd_tag <- "th~ td+ td"


  #Scrape the data items
  water.system.name <- the_website%>% html_nodes(the_water.system.name_tag)%>% html_text()
  PWSID <- the_website%>% html_nodes(the_pwsid_tag)%>% html_text()
  ownership <- the_website%>% html_nodes(the_ownership_tag)%>% html_text()
  max.withdrawals.mgd <- the_website%>% html_nodes(the_max_mgd_tag)%>% html_text()
  Year <- rep(the_year,times = 12)
  months <- c("Jan", "May", "Sep" ,"Feb" ,"Jun", "Oct" ,"Mar", "Jul", "Nov", "Apr" ,"Aug", "Dec")

  #Convert to a data frame
  df_withdrawls<-data.frame("Year"= rep(the_year,times = 12),
                            "Max_withdrawals_mgd" = as.numeric(max.withdrawals.mgd),
                            the_pwsid)%>%
      mutate( #PWSID = !!PWSID,
             Water_system_name = !!water.system.name,
             Ownership= !!ownership,
             Month = !!month,
             Date=my(paste(month,"-",Year)))

  #Return the data frame
  return(df_withdrawls)
}
```

```
#Run the function
the_df <- scrape.it(2022,'03-32-010')
#view(the_df)
```

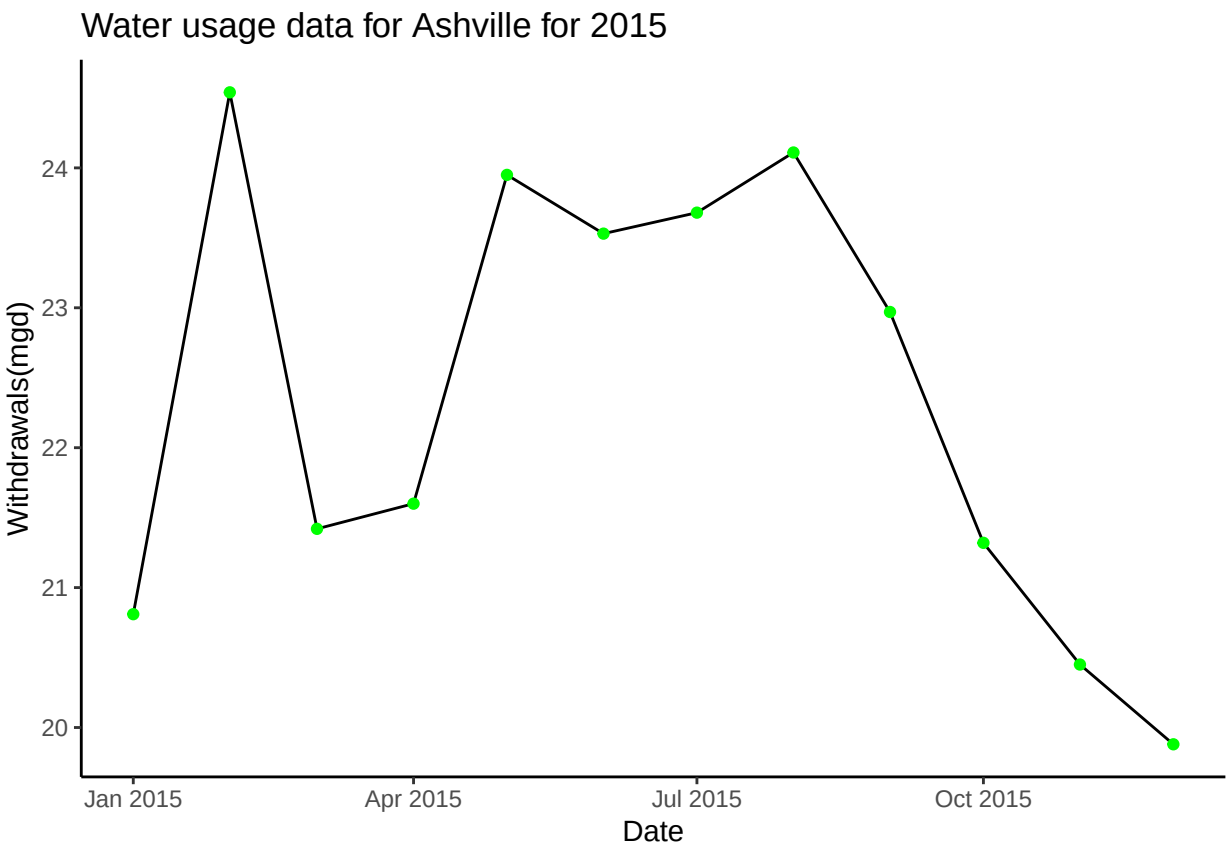
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID=‘03-32-010’) for each month in 2015

```
#7
the_pwsid <-'03-32-010'
the_year <-2015
the_df_2 <- scrape.it(the_year, the_pwsid)

ggplot(the_df_2, aes(x=Date, y=Max_withdrawals_mgd))+
       geom_line()+
         geom_point(color="pink")+
         labs(title = paste("2015 Water usage data for", water.system.name),
              y= "withdrawl (mgd)",
              x= "Date") +
  theme_classic()
```



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
the_pwsid <- '01-11-010'
the_year <- 2015
asheville <- scrape.it(the_year, the_pwsid)
#plotting the results
ggplot(asheville,aes(x=Date,y=Max_withdrawals_mgd))+
  geom_line()+
  geom_point(color="green")+
  labs(x="Date",
  y="Withdrawals(mgd)",
  title="Water usage data for Ashville for 2015") +
  theme_classic()
```



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').

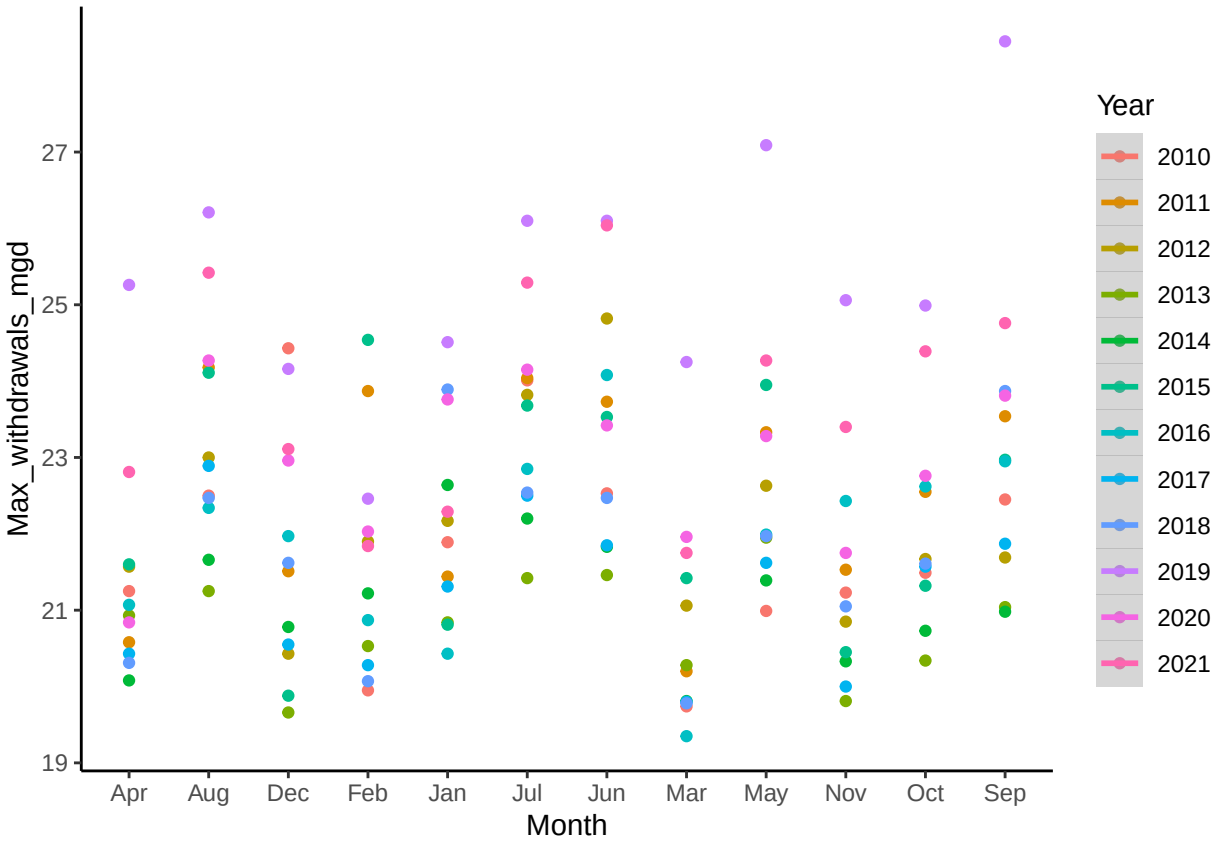
   TIP: See Section 3.2 in the "09_Data_Scraping.Rmd" where we apply "map2()" to iteratively run a function over two inputs. Pipe the output of the map2() function to `bindrows()` to combine the dataframes into a single one.

```
the_pwsid <- '01-11-010'
the_year <- 2015

the_website <- read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?', 'pwsid=',the_pwsid

  the_water.system.name_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_max_mgd_tag <- "th~ td+ td"


  #Scrape the data items
  water.system.name <- the_website%>% html_nodes(the_water.system.name_tag)%>% html_text()
  PWSID <- the_website%>% html_nodes(the_pwsid_tag)%>% html_text()
  ownership <- the_website%>% html_nodes(the_ownership_tag)%>% html_text()
  max.withdrawals.mgd <- the_website%>% html_nodes(the_max_mgd_tag)%>% html_text()
  Year <- rep(the_year,times = 12)
  months <- c("Jan", "May", "Sep" ,"Feb" ,"Jun", "Oct" ,"Mar", "Jul", "Nov", "Apr" ,"Aug", "Dec")
```

```
  #Convert to a data frame
  df_withdrawls<-data.frame("Year"= rep(the_year, times = 12),
                            "Max_withdrawals_mgd" = as.numeric(max.withdrawals.mgd),
                            the_pwsid)%>%
      mutate( #PWSID = !!PWSID,
             Water_system_name = !!water.system.name,
             Ownership= !!ownership,
             Month = !!month,
             Date=my(paste(month,"-",Year)))
```

```
#9
years <- seq.int(from = 2010, to = 2021, by = 1)
location_pwsid <- rep("01-11-010", 12)

asheville_map <- map2(years, location_pwsid, scrape.it) %>%
  bind_rows() %>%
  mutate(Date = ymd(Date)) %>%
  arrange(Date, Month) %>%
  mutate(Year = as.factor(Year))

ggplot(data = asheville_map ,aes(x = Month, y = Max_withdrawals_mgd, colour = Year)) +
  geom_point() +
  geom_smooth(method = "loess") +
  theme_classic()
```

```
## `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?