Assignment 10: Data Scraping

Aye Nyein Thu

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on data scraping.

Directions

1. Rename this file `<FirstLast>_A10_DataScraping.Rmd` (replacing `<sFirstLast>` 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 Set up
library(tidyverse)
library(lubridate)
library(rvest)
library(purrr)
library(here)
here()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

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

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham's 2023 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=2023

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

```
#2 Indicate the website as the URL to be scraped
Durham_LWSP <- read_html (
  "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2023")
Durham_LWSP
```

```
## {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)".

```
#3a Scraping 4 values and assigning them to separate variables
Water_system <- Durham_LWSP %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text
Water_system
```

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

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

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

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

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

```
Max_DayUse <- Durham_LWSP %>%
  html_nodes("th~ td+ td") %>%
  html_text
Max_DayUse
```

```
##  [1] "28.9000" "33.3000" "43.7000" "30.0000" "40.0000" "37.2300" "34.2000"
##  [8] "44.9000" "40.3500" "30.9000" "56.7000" "33.3000"
```

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.)

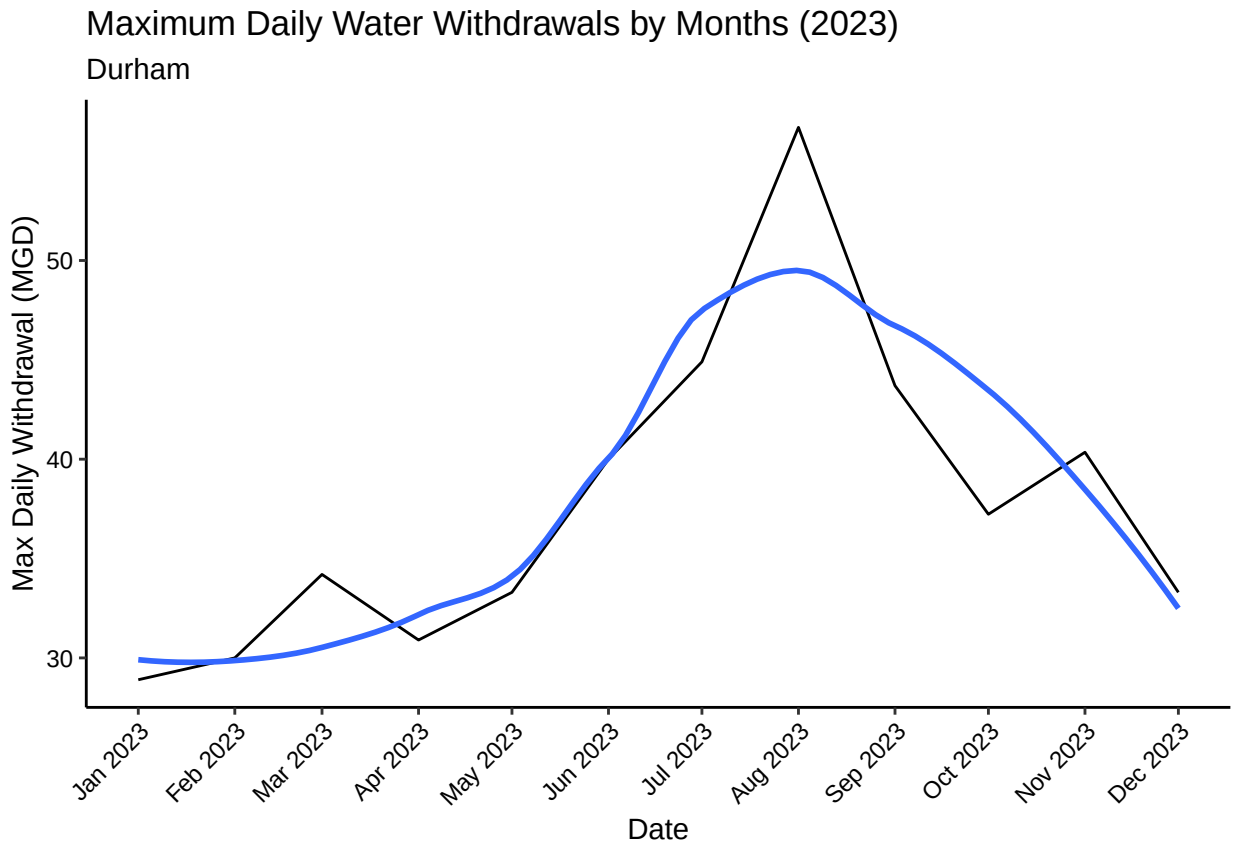
   TIP: Use `rep()` to repeat a value when creating a dataframe.

   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 maximum daily withdrawals across the months for 2023, making sure, the months are presented in proper sequence.

```
#4 Convert scraped data into a dataframe
Durham_2023 <-  data.frame(
  "Year" = rep(2023, 12),
  "Max_DayUse" = as.numeric(Max_DayUse)
) %>%
  mutate(Month=factor(
    c("Jan", "May", "Sep", "Feb","Jun","Oct","Mar",
      "Jul","Nov","Apr","Aug","Dec"),
                      levels=month.abb),
         Ownership=!!Ownership,
         PWSID=!!PWSID,
         Date = my(paste(Month,"-",Year)))

#5 Line Plot of the maximum daily withdrawls in 2023
ggplot(data = Durham_2023,
       aes(x=Date, y=Max_DayUse)) +
  geom_line() +
  geom_smooth(method="loess", se=FALSE) +
  labs(title="Maximum Daily Water Withdrawals by Months (2023)",
       subtitle = "Durham",
       y="Max Daily Withdrawal (MGD)",
```

```
      x="Date") +
  scale_x_date(date_labels = "%b %Y", date_breaks = "1 month") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



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, returning a dataframe. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

```
#6a. Assign pwsid, year and base url
the_base_url <- "https://www.ncwater.org/WUDC/app/LWSP/report.php"
the_pwsid <- "03-32-010"
the_year <- 2023
the_scrape_url <- paste0(the_base_url,"?pwsid=", the_pwsid,"&year=",the_year)
print(the_scrape_url)
```

```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2023"
```

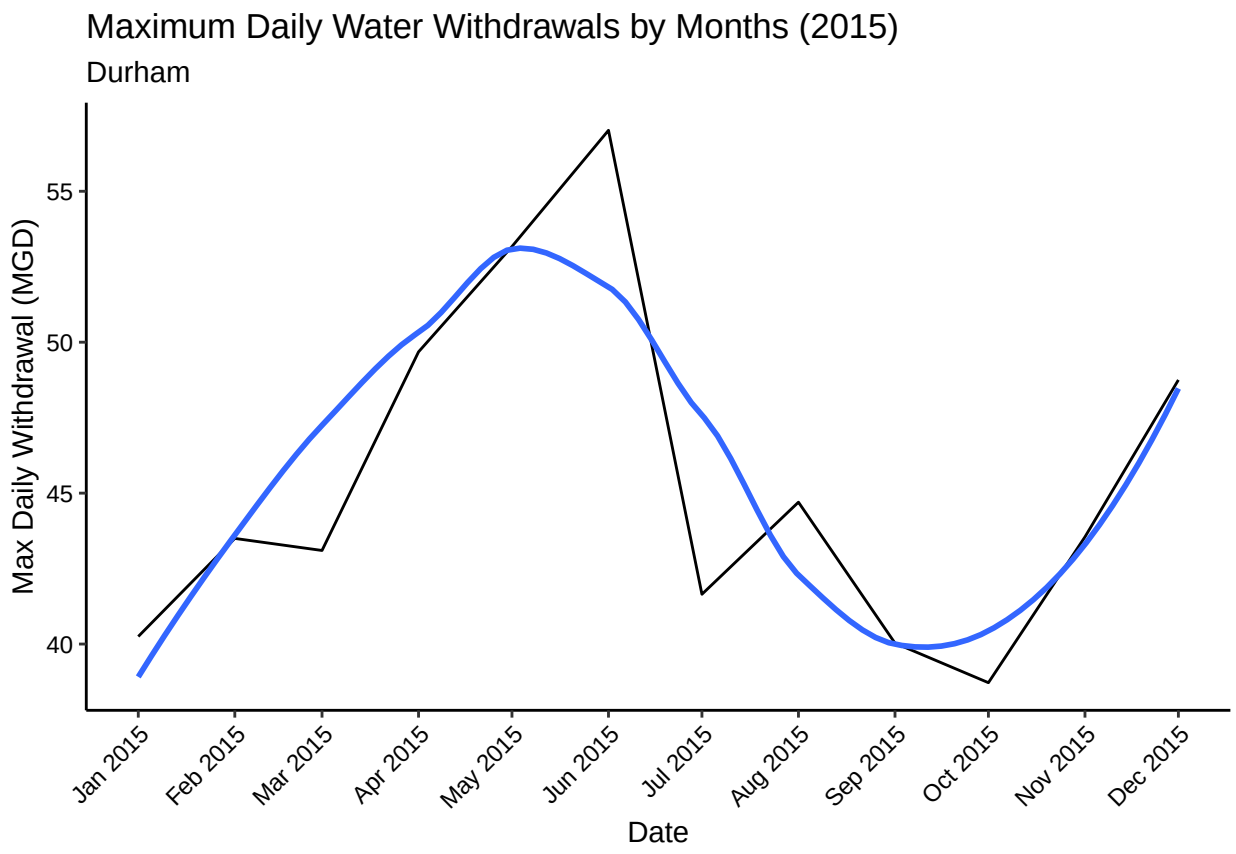
```
#6b. Create a scraping function
LWSP.scrape <- function(the_pwsid,the_year){

  #Retrieve the website
  LWSP_Website <- read_html(
    paste0("https://www.ncwater.org/WUDC/app/LWSP/report.php","?pwsid=",
```

```
          the_pwsid,"&year=",the_year))

  #Set the element address variables
  Water_system_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)"
  Max_DayUse_tag <- "th~ td+ td"

  #Scrape the data items
  Water_scr <- LWSP_Website %>% html_nodes(Water_system_tag) %>% html_text()
  PWSID_scr <- LWSP_Website %>% html_nodes(PWSID_tag) %>% html_text()
  Ownership_scr <- LWSP_Website %>% html_nodes(Ownership_tag) %>% html_text()
  Max_DayUse_scr <- LWSP_Website %>% html_nodes(Max_DayUse_tag) %>% html_text()

  #Convert to a dataframe
  LWSP_Website.scrape <-  data.frame(
  "Year" = rep(the_year, 12),
  "Max_DayUse" = as.numeric(Max_DayUse_scr)
) %>%
  mutate(
    Month=factor(
    c("Jan", "May", "Sep", "Feb","Jun","Oct","Mar",
      "Jul","Nov","Apr","Aug","Dec"),
                      levels=month.abb),
    Ownership=!!Ownership_scr,
         PWSID=!!PWSID_scr,
         Date = my(paste(Month,"-",Year)))

  #Pause for a moment as per the scraping etiquette
  Sys.sleep(1)

  #Return the dataframe
  return(LWSP_Website.scrape)
}
```

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

```
#7a Fetch Durham and 2015 values from the function
Durham_2015 <- LWSP.scrape("03-32-010",2015)
view(Durham_2015)


#7b Plot Max Daily Withdrawals
ggplot(data = Durham_2015,
       aes(x=Date, y=Max_DayUse)) +
  geom_line() +
  geom_smooth(method="loess", se=FALSE) +
  labs(title="Maximum Daily Water Withdrawals by Months (2015)",
       subtitle = "Durham",
       y="Max Daily Withdrawal (MGD)",
       x="Date") +
  scale_x_date(date_labels = "%b %Y", date_breaks = "1 month") +
```

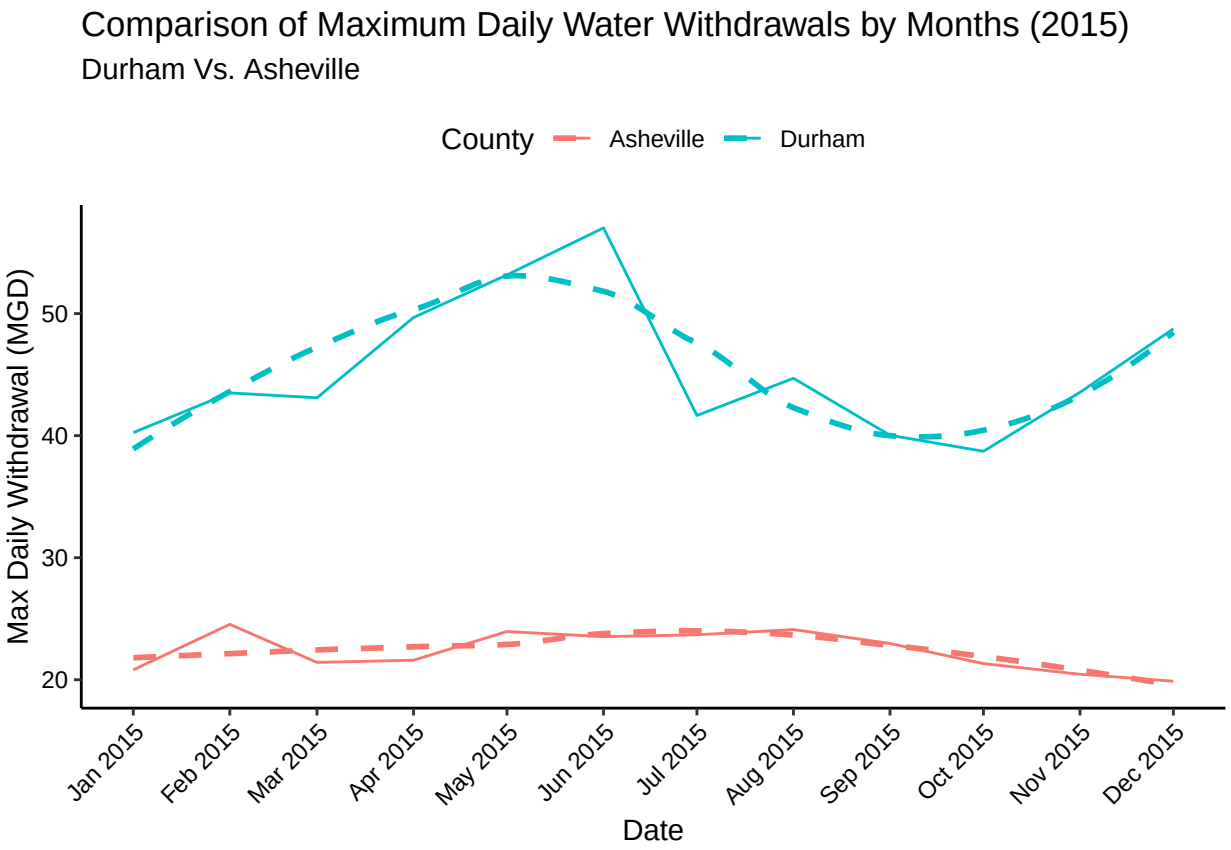
```
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



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.

```
#8a Fetch Asheville and 2015 values from the function
Asheville_2015 <- LWSP.scrape("01-11-010",2015)
view(Asheville_2015)

#8b Combine Asheville and Durham Data
Durham.Ashville_2015 <- bind_rows(Durham_2015,Asheville_2015) %>%
  mutate(County=case_when(
    PWSID == "03-32-010" ~ "Durham",
    PWSID == "01-11-010" ~ "Asheville"))

#8c Plot
ggplot(Durham.Ashville_2015, aes(x = Date, y = Max_DayUse, color = County)) +
  geom_line() +
  geom_smooth(method = "loess", se = FALSE, linetype = "dashed") +
  labs(title = "Comparison of Maximum Daily Water Withdrawals by Months (2015)",
       subtitle = "Durham Vs. Asheville",
       y = "Max Daily Withdrawal (MGD)",
       x = "Date",
       color = "County") +
```

```
  scale_x_date(date_labels = "%b %Y", date_breaks = "1 month") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

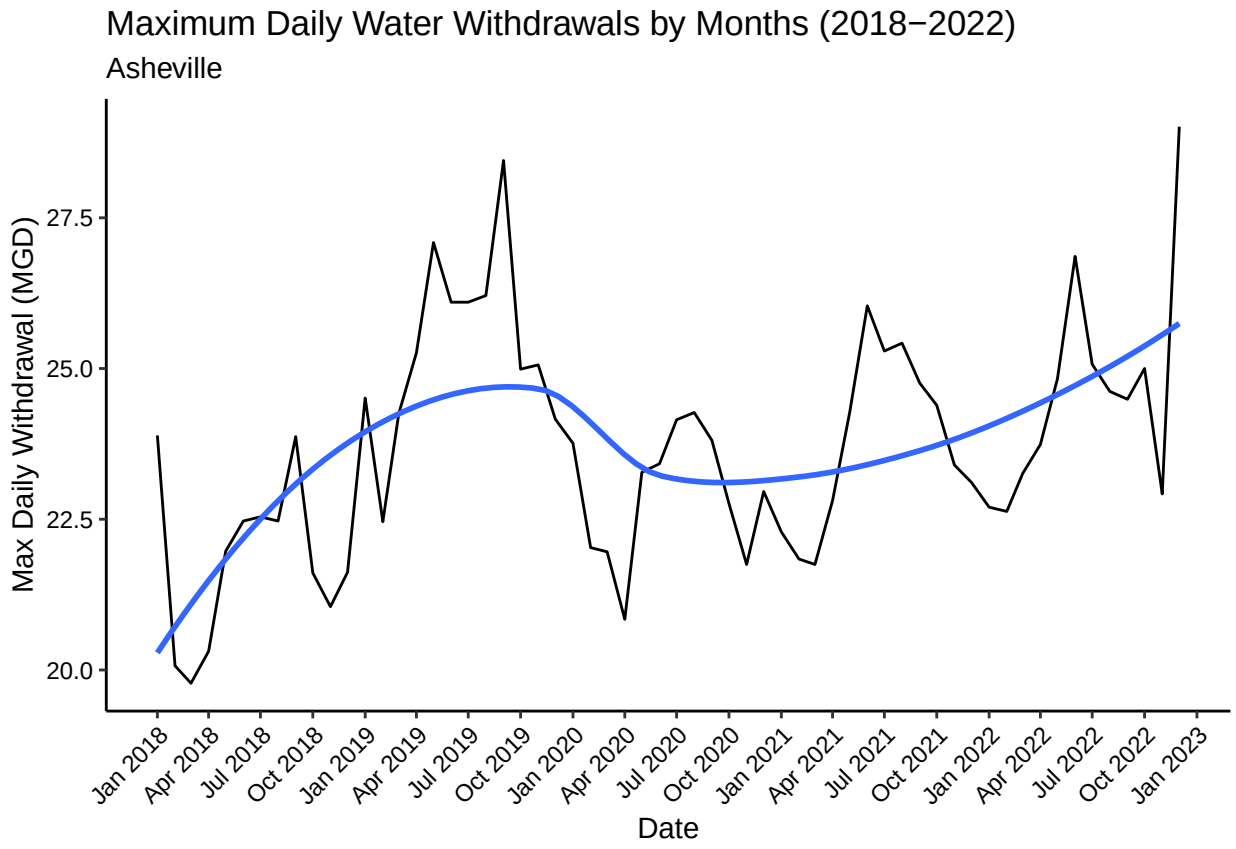


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2018 thru 2022.Add a smoothed line to the plot (method = 'loess').

   TIP: See Section 3.2 in the "10_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.

```
#9a Fetch Asheville and 2018 through 2022 values from the functio
years <- 2018:2022

Asheville_18.22 <- map2_dfr(
  rep("01-11-010", length(years)),
  years,
  ~LWSP.scrape(.x, .y)
)

#9b Plot
ggplot(Asheville_18.22, aes(x = Date, y = Max_DayUse)) +
  geom_line() +
  geom_smooth(method = "loess", se = FALSE) +
  labs(
```

```
    title = "Maximum Daily Water Withdrawals by Months (2018-2022)",
    subtitle = "Asheville",
    y = "Max Daily Withdrawal (MGD)",
    x = "Date"
  ) +
  scale_x_date(date_labels = "%b %Y", date_breaks = "3 month") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time? > Answer: Just by looking at the plot and without any additional data analysis, it could be estimated that Asheville has an increasing water usage trend over time. The usage was increasing rapidly during 2018 and 2019. However, there was a decrease in 2020 - the temporary closedown of businesses, offices and restaurants might contribute to the decreasing trend. Since the beginning of 2021, the water usage trend has been surging again.