Assignment 10: Data Scraping

Jazmine Pritchett

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(here)
install.packages("rvest")
library(rvest)
here()
```

```
## [1] "/home/guest/EDE_Fall2023"
```

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

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

```
## {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)".

```
#3
water_system <- website %>%
  html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>%
  html_text()

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

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

MGD <- website %>%
  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.)

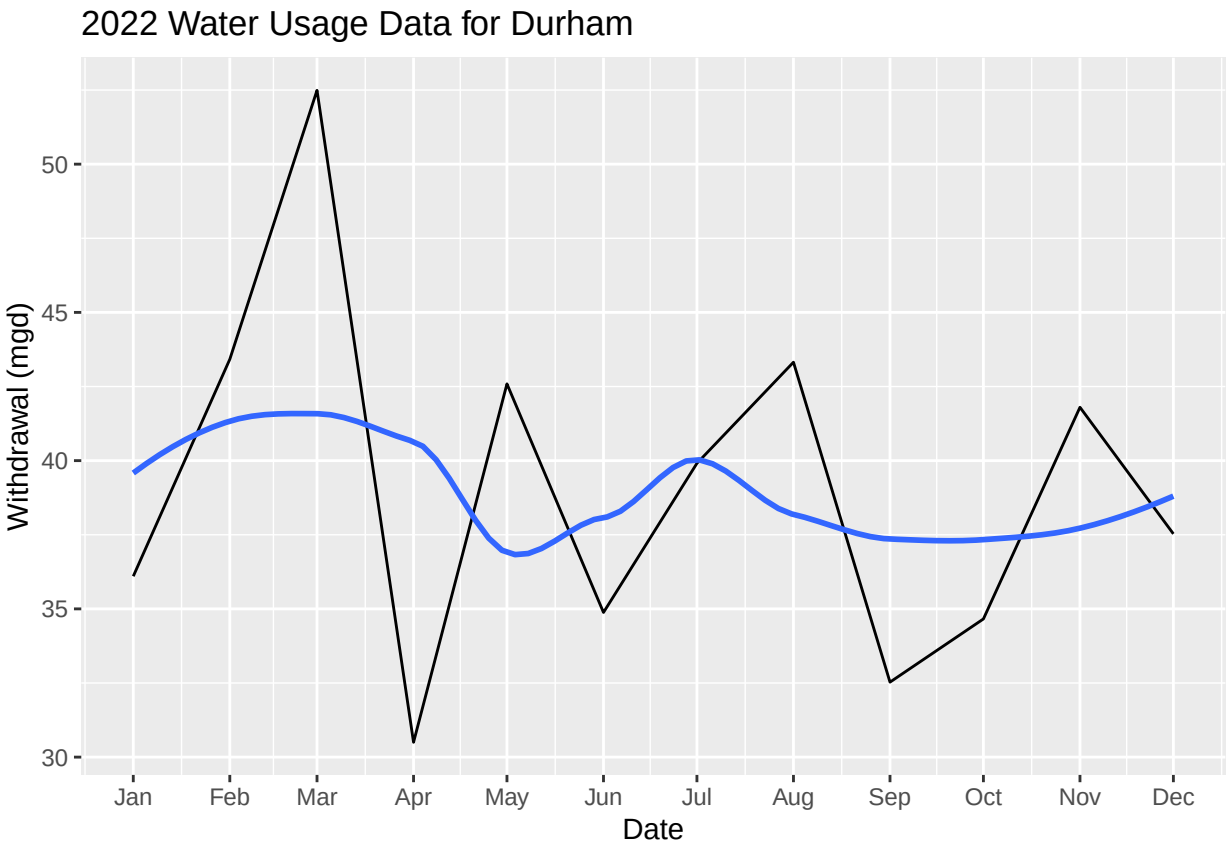
> TIP: Use `rep()` to repeat a value when creating a dataframe.

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 maximum daily withdrawals across the months for 2022

```
#4
df_system <- data.frame(water_system = rep(water_system, 12),PWSID = rep(PWSID, 12),
ownership = rep(ownership, 12), MGD = as.numeric(MGD, 12),
month = rep(c("Jan", "Feb", "Mar", "Apr", "May",
"Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec")))


df_system$Date <- as.Date(paste0("2022-", formatC(as.numeric(factor(df_system$month,
levels = c("Jan", "Feb", "Mar", "Apr", "May",
"Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"))),
width = 2, flag = "0"), "-01"))

#5
ggplot(df_system,aes(x=Date,y=MGD)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2022 Water Usage Data for",water_system),
       y="Withdrawal (mgd)",
       x="Date")+
  scale_x_date(date_breaks = "1 month", date_labels = "%b")
```

```
## `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 (pwsid) scraped**.

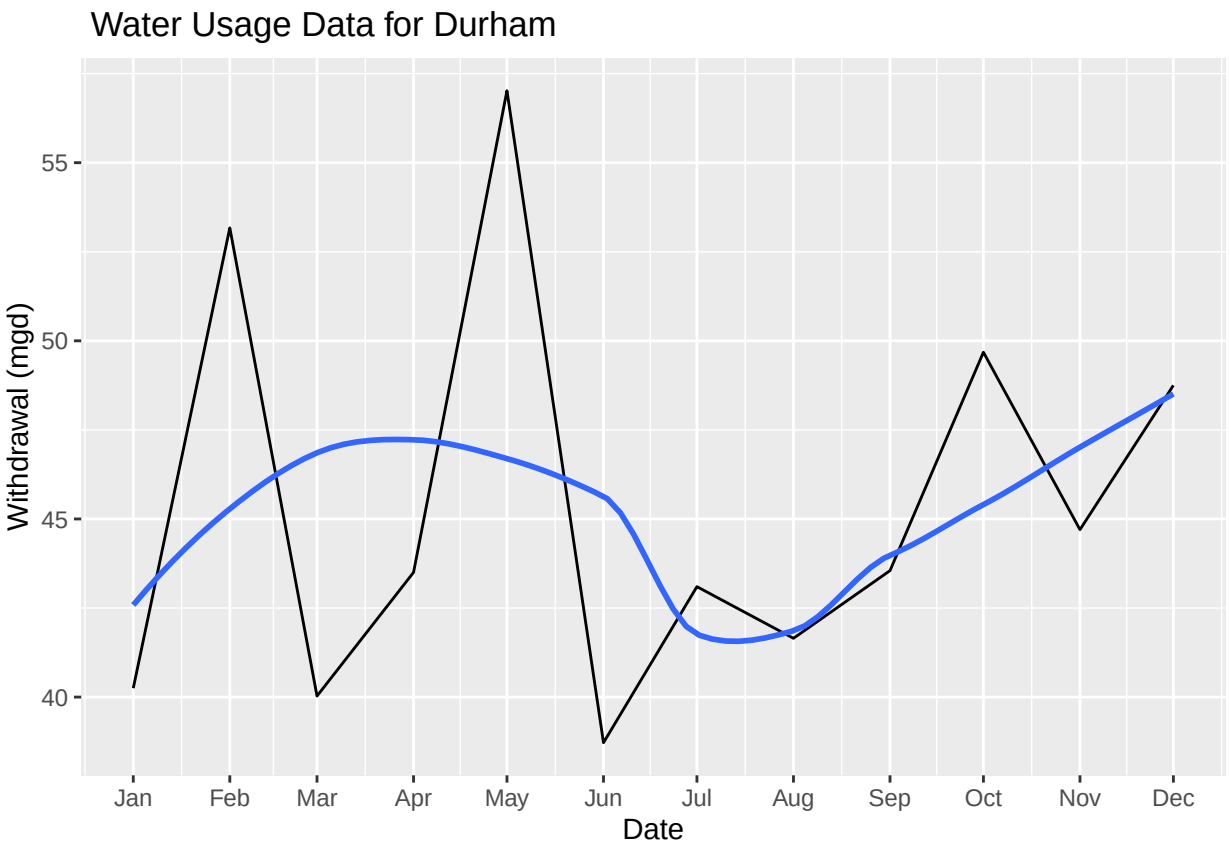
```
#6.
scrape_water_data <- function(pwsid, year) {
  url <- paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=', pwsid, '&year=',
year)
  website <- read_html(url)

  water_system <- website %>% html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>%
  html_text()
  PWSID <- website %>% html_nodes('td tr:nth-child(1) td:nth-child(5)') %>% html_text()
  ownership <- website %>% html_nodes('div+ table tr:nth-child(2) td:nth-child(4)') %>%
  html_text()
  MGD <- website %>% html_nodes('th~ td+ td') %>% html_text()

  df_system <- data.frame(water_system = rep(water_system, 12),
                          PWSID = rep(PWSID, 12),
                          ownership = rep(ownership, 12),
                          MGD = as.numeric(MGD, 12),
                          month = rep(c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep",
"Oct", "Nov", "Dec"), stringsAsFactors = FALSE))

  df_system$Date <- as.Date(paste0(year, "-", formatC(as.numeric(factor(df_system$month,
```

```
levels = c("Jan", "Feb", "Mar", "Apr", "May",
"Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"))),
width = 2, flag = "0"), "-01"))

ggplot(df_system, aes(x = Date, y = MGD)) +
  geom_line() +
  geom_smooth(method = "loess", se = FALSE) +
  labs(title = paste(unique(df_system$year), "Water Usage Data for",
      unique(df_system$water_system)),
       y = "Withdrawal (mgd)",
       x = "Date") +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")
}
```

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

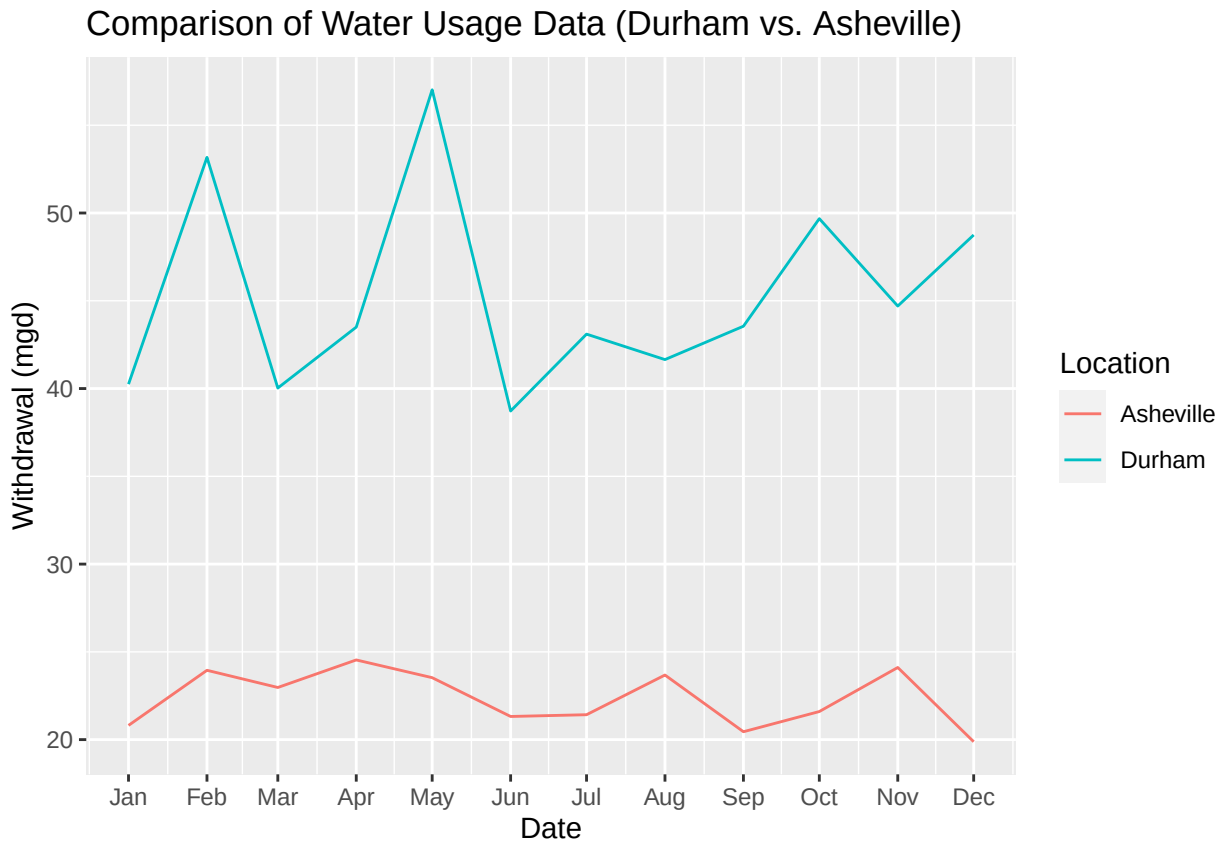
```
#7

scrape_water_data("03-32-010", 2015)
```

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



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

durham_data <- scrape_water_data("03-32-010", 2015)
asheville_data <- scrape_water_data("01-11-010", 2015)

# Combine data frames for Durham and Asheville
combined_data <- bind_rows(asheville_data$data, durham_data$data, .id = "location")

# Create a comparison plot
ggplot(combined_data, aes(x = Date, y = MGD, color = water_system)) +
  geom_line() +
  labs(title = "Comparison of Water Usage Data (Durham vs. Asheville)",
       y = "Withdrawal (mgd)",
       x = "Date",
       color = "Location")+
  scale_x_date(date_breaks = "1 month", date_labels = "%b")
```

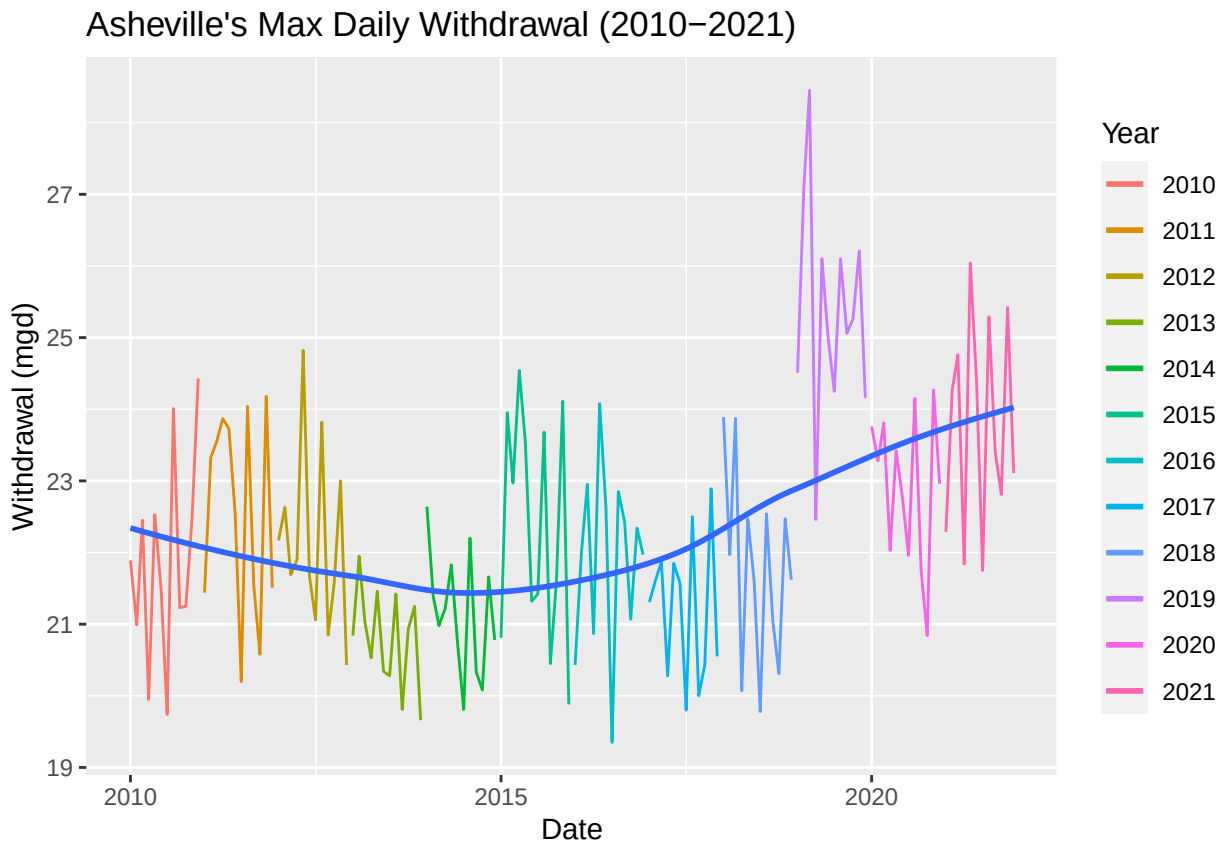


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2021.Add a smoothed line to the plot (method = 'loess').

   TIP: See Section 3.2 in the "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.

```
#9


asheville_facility <- lapply(2010:2021, function(year) {
  scrape_water_data("01-11-010", year)$data
})

asheville_combined <- bind_rows(asheville_facility, .id = "Year")

ggplot(asheville_combined, aes(x = Date, y = MGD, color = as.factor(year(Date)))) +
  geom_line() +
  geom_smooth(method = "loess", se = FALSE, aes(group = 1)) +
  labs(title = "Asheville's Max Daily Withdrawal (2010-2021)",
       y = "Withdrawal (mgd)",
       x = "Date") +
  scale_color_discrete(name = "Year")
```



Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time? > Answer: > Over time, the trend for water usage has been increasing since around 2015 after it was decreasing before then.