Assignment 10: Data Scraping

Iddrisu Sharu Deen

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(rvest); library(here); library(lubridate); library(purrr)

here()
```

```
## [1] "/home/guest/EDA_Spring2024"
```

```
mytheme <- theme_minimal() +
  theme(
    text = element_text(size = 12),
    plot.title = element_text(face = "bold", size = 14),
    plot.subtitle = element_text(face = "italic", size = 12),
    axis.title = element_text(face = "bold"),
    axis.text = element_text(size = 8.5),
    legend.position = "bottom",
    legend.title = element_blank())

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
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)".

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

```
## [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"
```

```
max_day_use_per_month <- webpage %>%
  html_nodes("th~ td+ td") %>%
  html_text()
max_day_use_per_month
```

```
##  [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

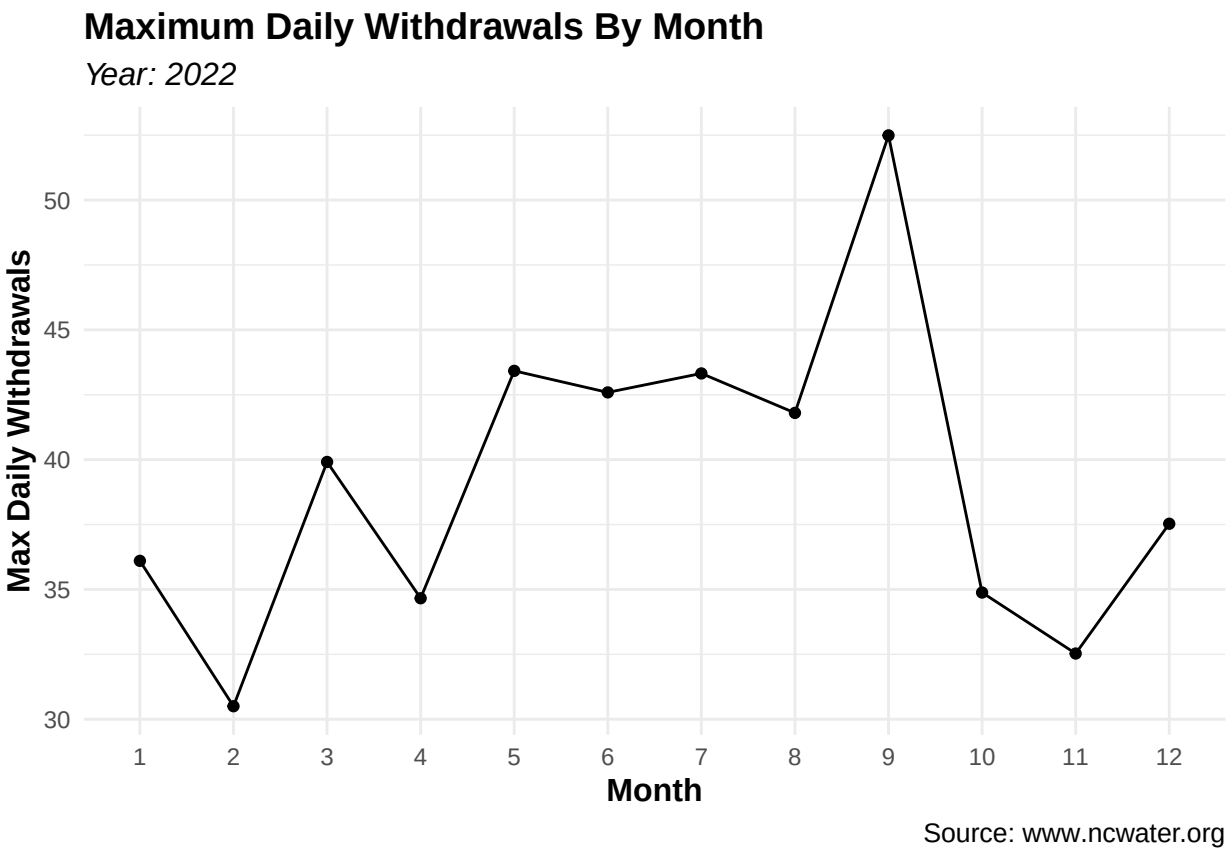
column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

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 maximum daily withdrawals across the months for 2022

```
#4
durham22.df <- data.frame("Month" = as.factor(c(1,5,9,2,6,10,3,7,11,4,8,12)),
                          "Year" = rep(2022, 12),
                          "Water System" = water_system,
                          "Ownership" = ownership,
                          "Max Daily Usage Per Month" = as.numeric(max_day_use_per_month)
                          )

durham22.df <- arrange(durham22.df, by_group = Month)

#5
max_daily_withdrawals.plt <- ggplot(durham22.df, aes(x=Month, y=Max.Daily.Usage.Per.Month), group = 1) +
  geom_line(group=1) +
  geom_point()+
  labs(title = "Maximum Daily Withdrawals By Month",
       subtitle = "Year: 2022",
       caption = "Source: www.ncwater.org",
       x = "Month",
       y = "Max Daily WIthdrawals")
max_daily_withdrawals.plt
```



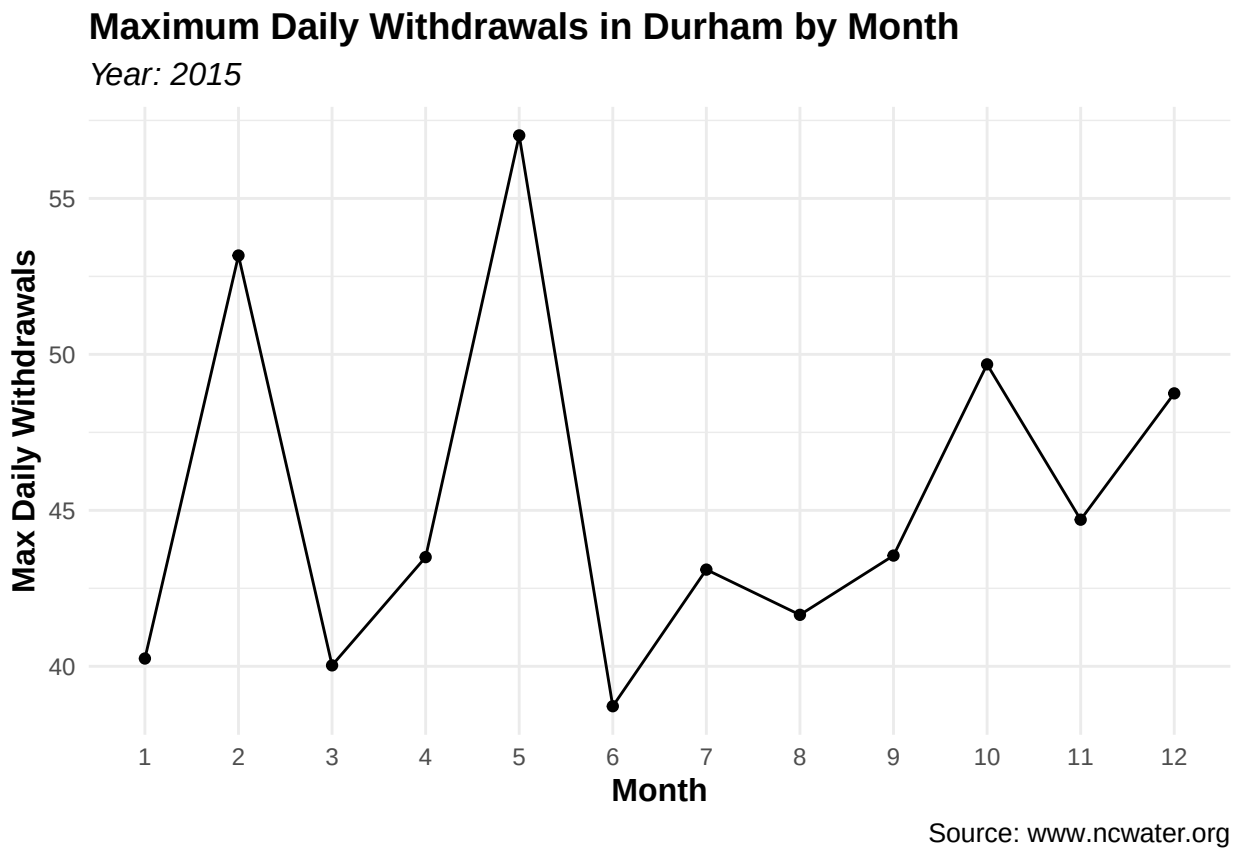
6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

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

  # Set variables
  water_system.tag <- "div+ table tr:nth-child(1) td:nth-child(2)"
  pwsid.scrape.tag <- "td tr:nth-child(1) td:nth-child(5)"
  ownership.tag <- "div+ table tr:nth-child(2) td:nth-child(4)"
  usage.tag <- "th~ td+ td"

  # Scrape data
  water_system.scrape <- site.url %>% html_nodes(water_system.tag) %>% html_text()
  pwsid.scrape <- site.url %>% html_nodes(pwsid.scrape.tag) %>% html_text()
  ownership.scrape <- site.url %>% html_nodes(ownership.tag) %>% html_text()
  usage.scrape <- site.url %>% html_nodes(usage.tag) %>% html_text()

  # Convert to dataframe
  withdrawals.df <- data.frame("Month" = factor(1:12),
                               "Year" = rep(year, 12),
                               "Max.Daily.Usage.Per.Month" = as.numeric(usage.scrape)) %>%
    mutate(Water.System = water_system.scrape,
```

```
            PWSID = pwsid.scrape,
            Ownership = ownership.scrape,
            Date = make_date(Year, Month, day = 1))

  # Pause
  Sys.sleep(1)

  # Output
  return(withdrawals.df)
}
```

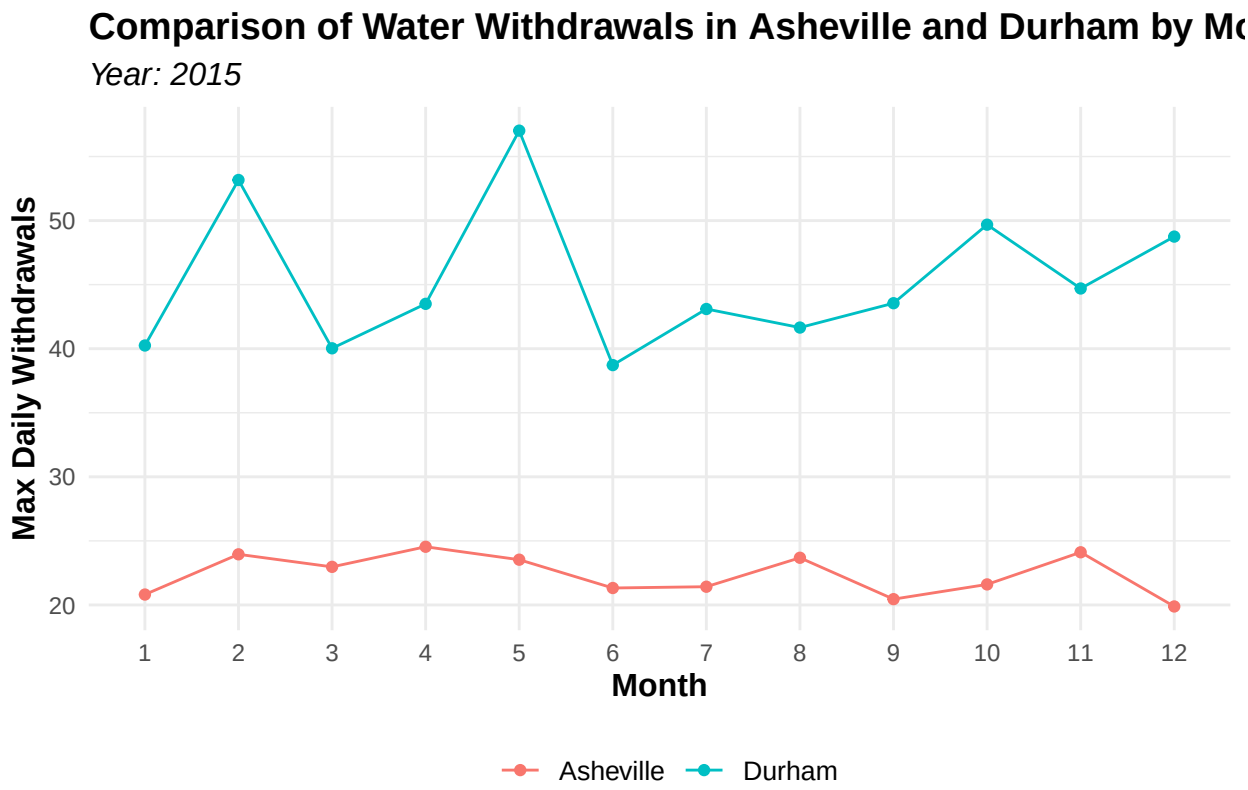
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID='03-32-010') for each month in 2015

```
#7
durham15.df <- scrape.function(2015, "03-32-010")

durham15.plt <- ggplot(durham15.df, aes(x = Month, y = Max.Daily.Usage.Per.Month)) +
  geom_point() +
  geom_line(group = 1) +
  labs(
    title = "Maximum Daily Withdrawals in Durham by Month",
    subtitle = "Year: 2015",
    caption = "Source: www.ncwater.org",
    x = "Month",
    y = "Max Daily Withdrawals"
  )
durham15.plt
```



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.

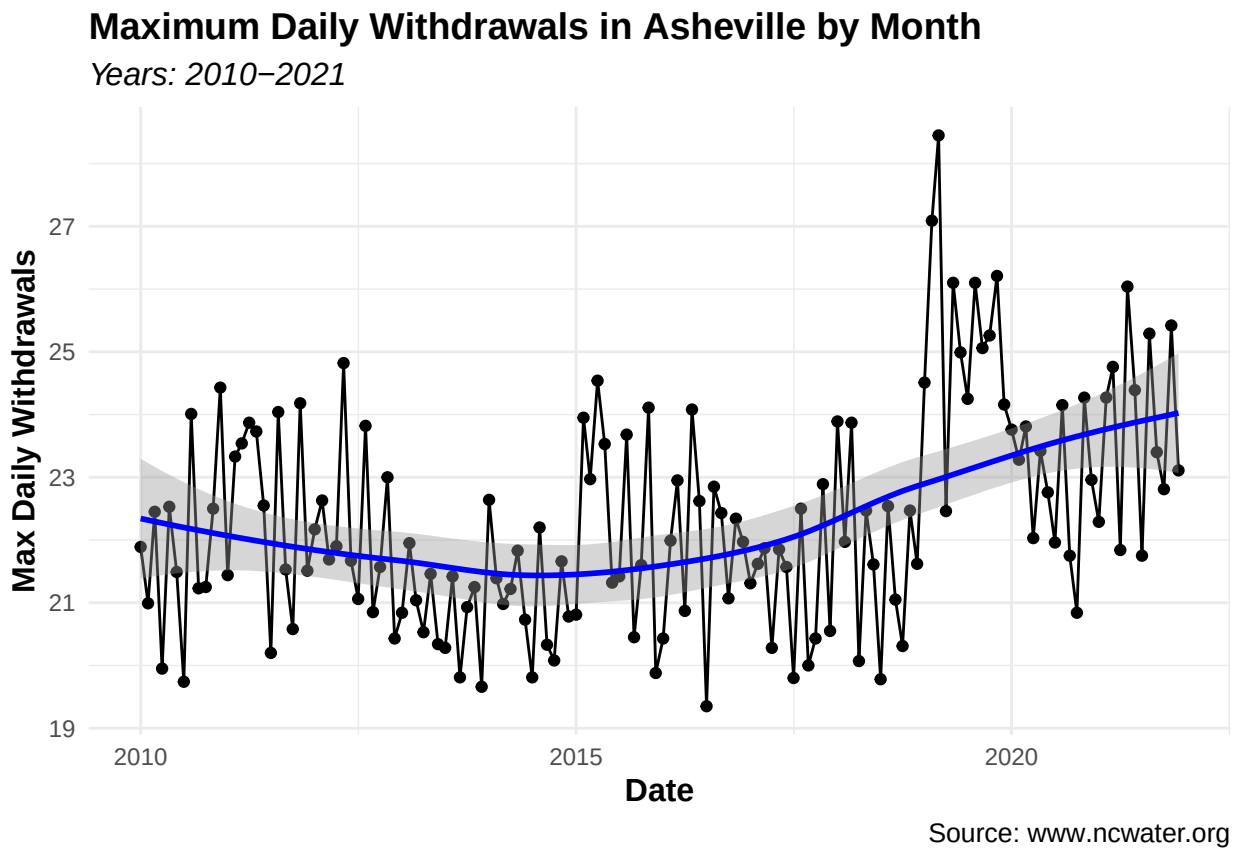
```
#8
asheville15.df <- scrape.function(2015, "01-11-010")

combined_data <- bind_rows(durham15.df, asheville15.df)

combined_plt <- ggplot(combined_data, aes(x = Month, y = Max.Daily.Usage.Per.Month, color = Water.Syste
  geom_point() +
  geom_line(aes(group = Water.System)) +
  labs(
    title = "Comparison of Water Withdrawals in Asheville and Durham by Month",
    subtitle = "Year: 2015",
    caption = "Source: www.ncwater.org",
    x = "Month",
    y = "Max Daily Withdrawals"
  )
combined_plt
```



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
years10_21 <- 2010:2021
asheville_pwsid = "01-11-010"

asheville10_21.df <- asheville10_21.df <- map2(years10_21, asheville_pwsid , scrape.function) %>%
  bind_rows()

asheville10_21.plt <- ggplot(asheville10_21.df, aes(x = Date, y = Max.Daily.Usage.Per.Month)) +
  geom_point() +
  geom_line() +
  geom_smooth(method = 'loess', se = TRUE, color = "blue") +
  labs(
    title = "Maximum Daily Withdrawals in Asheville by Month",
    subtitle = "Years: 2010-2021",
    caption = "Source: www.ncwater.org",
    x = "Date",
    y = "Max Daily Withdrawals"
  )
asheville10_21.plt
```



Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time? > Answer: There appears to be an overall slight positive trend in water usage in Asheville