Assignment 10: Data Scraping

Emily Kuhlmann

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)
library(lubridate)
library(rvest)

getwd()
```

```
## [1] "/Users/emilykuhlmann/Documents/Documents/Fall23/EDA/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
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)".

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

```
## [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 <- webpage %>%
  html_nodes("th~ td+ td , th~ td+ td") %>%
  html_text()
Max_Day_Use
```

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

df_dayuse <- data.frame("Month" = Months,
                        "Year" = rep(2022),
                        "Water_System" = water_sys,
                        "PWSID" = PWSID,
                        "Ownership" = Ownership,
                        "Max_Day_Use" = as.numeric(Max_Day_Use))

df_dayuse <- df_dayuse %>%
  mutate(Date = my(paste(Month, "-", Year)))

#5
MaxDailyWithdrawls <- ggplot(df_dayuse, aes(x = Date, y = Max_Day_Use)) +
  geom_line() +
  scale_x_date(date_breaks = "1 month", date_labels = "%b") +
  labs(y = "Max Daily Withdrawls (MGD)", x = "Month")
MaxDailyWithdrawls
```

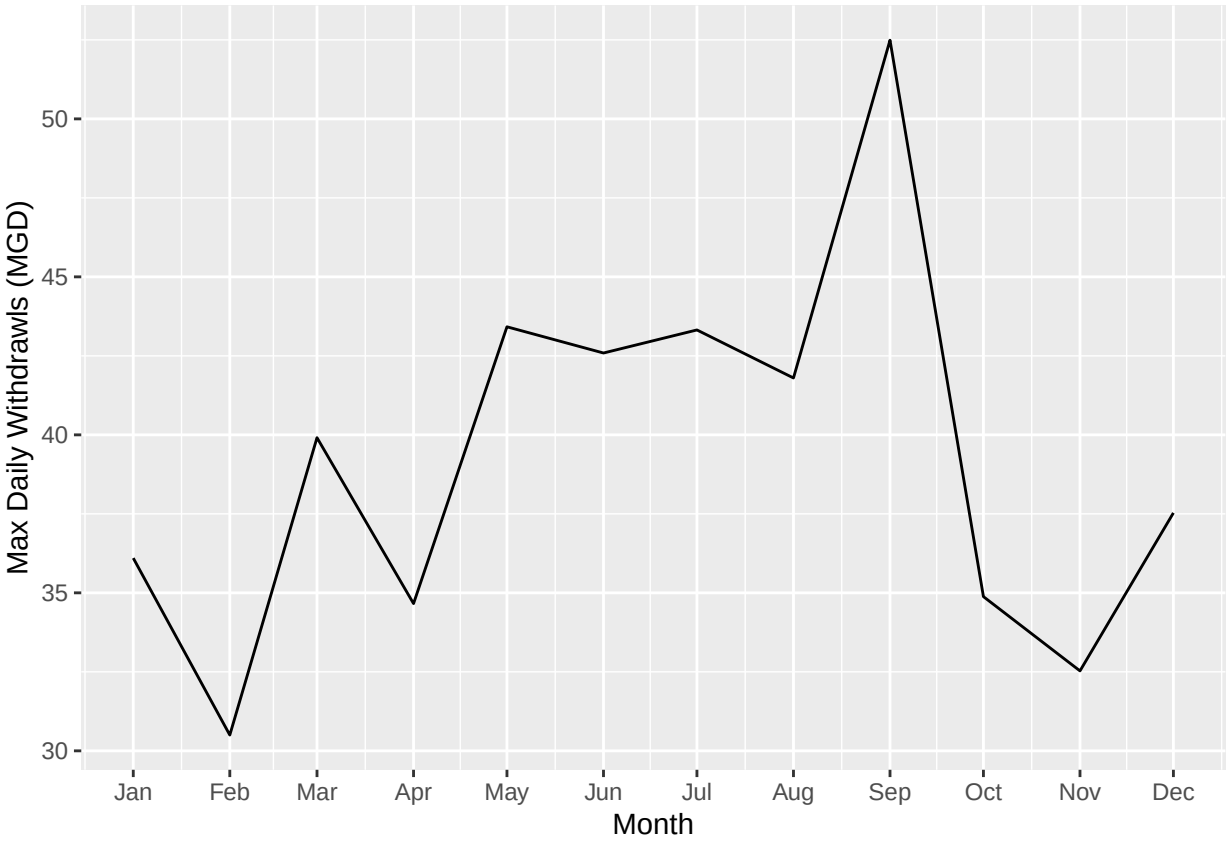

Figure 1: Maximum daily water withdrawls for Durham in 2022

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){
  #Get the proper url
  the_url <- ifelse(
    the_year=='2022' & the_PWSID =='03-32-010',
    'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2022',
    paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',
           the_PWSID,'&year=',the_year))

  #Fetch the website
  the_website <- read_html(the_url)

  #Scrape the data
  the_water_sys <- the_website %>%
    html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>% html_text()
  the_site_id <- the_website %>%
    html_nodes("td tr:nth-child(1) td:nth-child(5)") %>% html_text()
  the_ownership <- the_website %>%
    html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>% html_text()
  the_max_use <- the_website %>% html_nodes("th~ td+ td , th~ td+ td") %>%
  html_text()

  #Convert to dataframe
  df_dayuse <- data.frame("Month" = c("Jan", "May", "Sep", "Feb", "Jun", "Oct",
                                      "Mar", "Jul", "Nov", "Apr", "Aug", "Dec"),
                          "Year" = rep(the_year,12),
                          "Water_System" = the_water_sys,
                          "PWSID" = the_site_id,
                          "Ownership" = the_ownership,
                          "Max_Day_Use" = as.numeric(the_max_use)) %>%
    mutate(Date = my(paste(Month, "-", Year)))


  #Return the dataframe
  return(df_dayuse)
}
```

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_2015 <- scrape.it("2015", "03-32-010")

Drm2015 <- ggplot(Durham_2015, aes(x = Date, y = Max_Day_Use)) +
  geom_line() +
  scale_x_date(date_breaks = "1 month", date_labels = "%b") +
  labs(y = "Max Daily Withdrawls (MGD)", x = "Month")
Drm2015
```

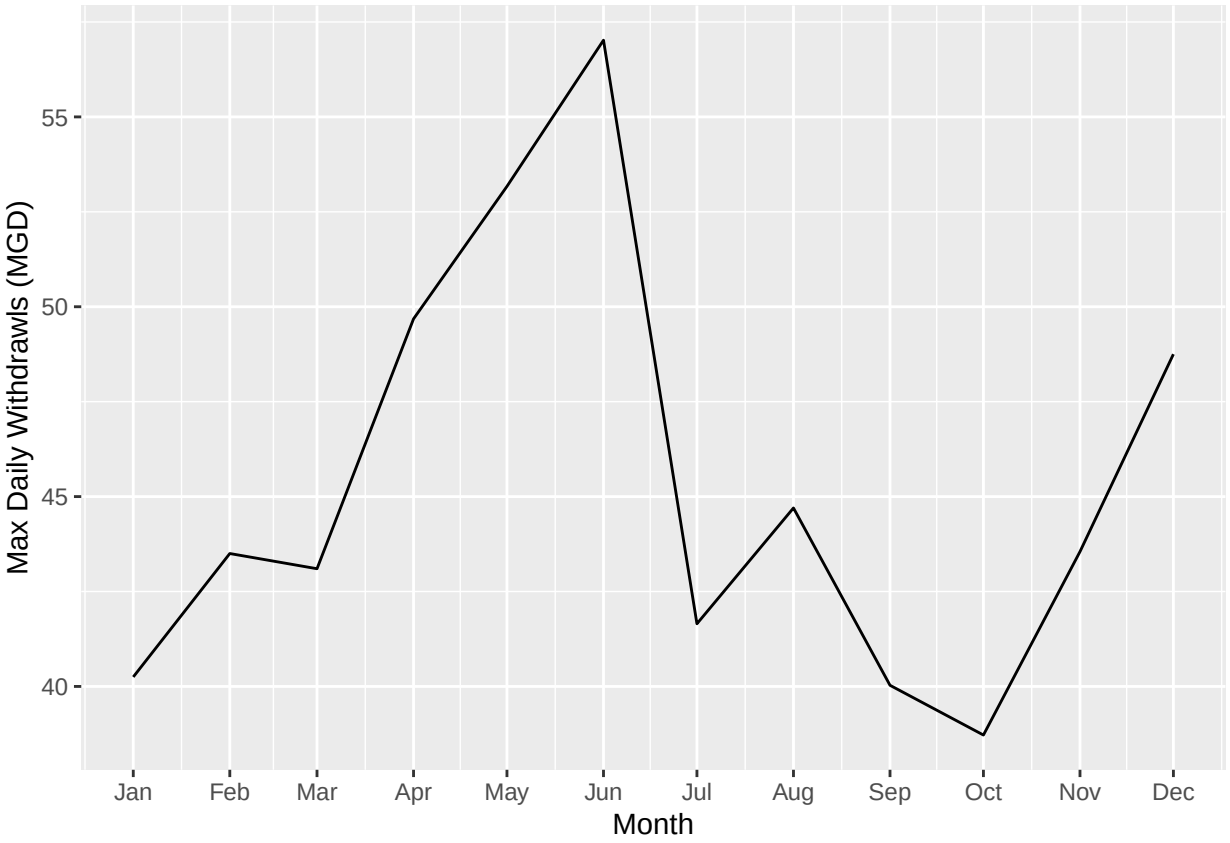

Figure 2: Maximum montly water withdrawls in Durham for 2015

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_2015 <- scrape.it("2015", "01-11-010")

Drm_Ashv_2015 <- ggplot() +
  geom_line(Durham_2015, mapping = aes(x = Date, y = Max_Day_Use,
                                       color = Water_System)) +
  geom_line(Asheville_2015, mapping = aes(x = Date, y = Max_Day_Use,
                                          color = Water_System)) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b") +
  labs(y = "Max Daily Withdrawls (MGD)", x = "Month", color = "Water System",
       title = "Maximum monthly water withdrawls, 2015")
Drm_Ashv_2015
```

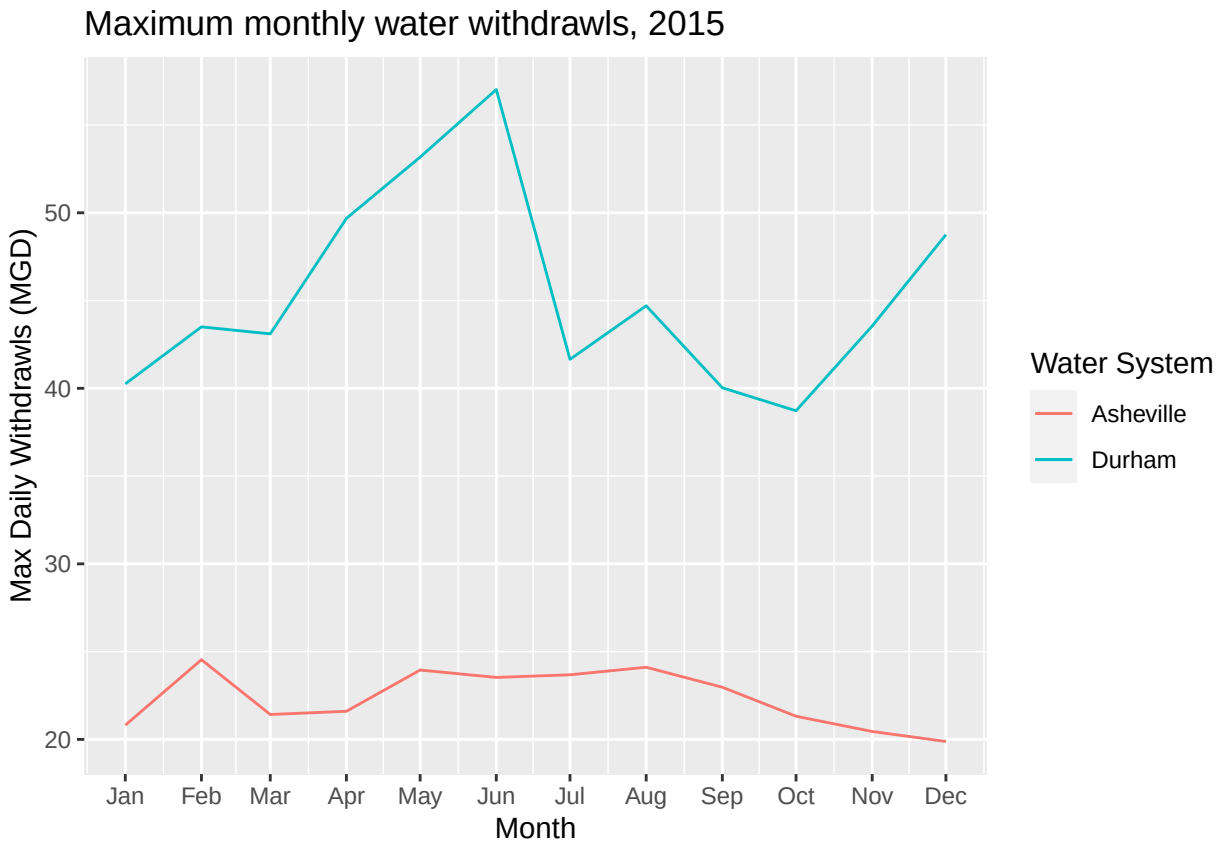

Figure 3: Maximum montly water withdrawls in Durham and Asheville for 2015

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.

```r
#9
the_years <- c(2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019,
               2020, 2021)

Ash_2010_21 <- map2(the_years, "01-11-010", scrape.it) %>%
  bind_rows()

Asheville_10_21 <- ggplot(Ash_2010_21, aes(x = Date, y = Max_Day_Use)) +
  geom_line() +
  geom_smooth(method = 'loess') +
  scale_x_date(date_breaks = "1 year", date_labels = "%Y") +
  labs(y = "Maximum daily withdrawls, MGD", x = "Year")
Asheville_10_21
```

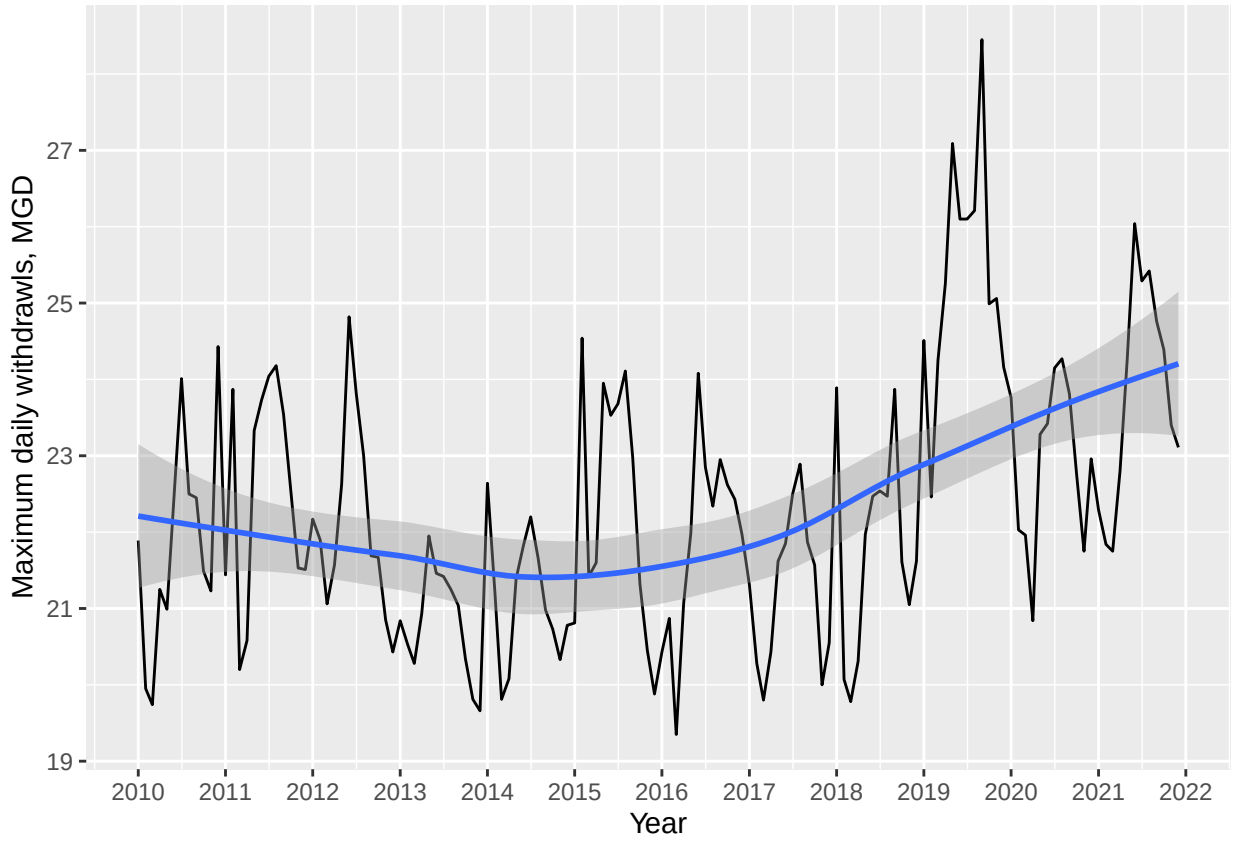

Figure 4: Maximum water withdrawls in Asheville from 2015 to 2021

Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time? > Answer:It appears that Asheville has an overall positive trend in water usage over this period of time, especially from 2015 to 2021. >