Assignment 10: Data Scraping

Jingze Dai

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(here)
library(tidyverse)
library(rvest)
getwd()
```

```
## [1] "/home/guest/ENV872/EDE_Fall2024"
```

```
here()
```

```
## [1] "/home/guest/ENV872/EDE_Fall2024"
```

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
url <-
  "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2023"
webpage <- read_html(url)
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
# creating tags
water_system_tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'
pswid_tag <- 'td tr:nth-child(1) td:nth-child(5)'
ownership_tag <- 'div+ table tr:nth-child(2) td:nth-child(4)'
day_use_tag <- 'th~ td+ td'

# data scraping
water_system_name <- webpage %>%
  html_nodes(water_system_tag) %>% html_text()
pswid <- webpage %>%
  html_nodes(pswid_tag) %>%  html_text()
ownership <- webpage %>%
  html_nodes(ownership_tag) %>% html_text()
day_use <- webpage %>%
  html_nodes(day_use_tag) %>% html_text()

# checking result
water_system_name
```

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

```
pswid
```

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

```
ownership
```

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

```
day_use
```

```
##  [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.

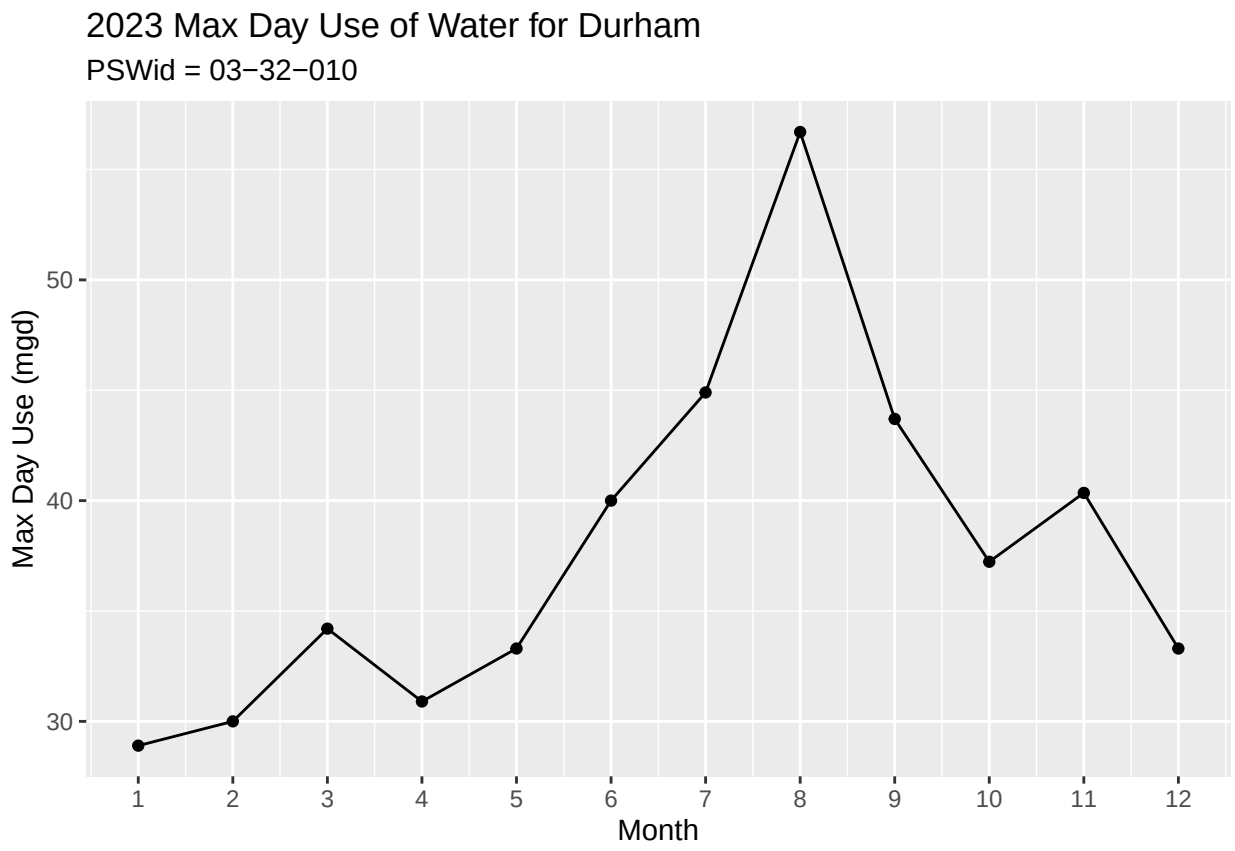
   NOTE: It's likely you won't be able to scrape the monthly widthrawal data in chronological order. You can overcome this by creating a month column manually assigning values in the order the data are scraped: "Jan", "May", "Sept", "Feb", etc... Or, you could scrape month values from the web page...

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
df_durham_LWSP <- data.frame("Month" = c(1,5,9,2,6,10,3,7,11,4,8,12),
                             "Year" = rep(2023),
                             "Maximum_Day_Use_mgd" = as.numeric(day_use)) %>%
  mutate(Water_system_name = !!water_system_name,
         PSWid = !!pswid,
         Ownership = !!ownership,
         Date = ym(paste(Year,"-",Month)))

# checking dataset
glimpse(df_durham_LWSP)
```

```
## Rows: 12
## Columns: 7
## $ Month               <dbl> 1, 5, 9, 2, 6, 10, 3, 7, 11, 4, 8, 12
## $ Year                <dbl> 2023, 2023, 2023, 2023, 2023, 2023, 2023, 2023, 20~
## $ Maximum_Day_Use_mgd <dbl> 28.90, 33.30, 43.70, 30.00, 40.00, 37.23, 34.20, 4~
## $ Water_system_name   <chr> "Durham", "Durham", "Durham", "Durham", "Durham", ~
## $ PSWid               <chr> "03-32-010", "03-32-010", "03-32-010", "03-32-010"~
## $ Ownership           <chr> "Municipality", "Municipality", "Municipality", "M~
## $ Date                <date> 2023-01-01, 2023-05-01, 2023-09-01, 2023-02-01, 20~
```

```r
#5
ggplot(df_durham_LWSP,aes(x=Month,y=Maximum_Day_Use_mgd)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks = 1:12) +
  labs(title = paste("2023 Max Day Use of Water for",water_system_name),
       subtitle = paste("PSWid =",pswid),
       y="Max Day Use (mgd)",
       x="Month")
```



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**.

```r
#6.
scrape_lwsp <- function(pwsid, year) {
  target_url <- paste0("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=",
                       pwsid, "&year=", year)
  target_webpage <- read_html(target_url)

  water_system_name <- target_webpage %>%
    html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>%
    html_text()
```

```
  pswid <- target_webpage %>%
    html_nodes('td tr:nth-child(1) td:nth-child(5)') %>%
    html_text()

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

  day_use <- target_webpage %>%
    html_nodes('th~ td+ td') %>%
    html_text() %>%
    as.numeric()

  df.selected <- data.frame(
    "Month" = c(1,5,9,2,6,10,3,7,11,4,8,12),
    "Year" = rep(year),
    "Maximum_Day_Use_mgd" = day_use
  ) %>%
    mutate(
      Water_system_name = water_system_name,
      PSWid = pswid,
      Ownership = ownership,
      Date = ym(paste(Year,"-",Month))
    )
  return(df.selected)
}
```

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

```
#7
df_durham_2015 <- scrape_lwsp("03-32-010", 2015)
# checking result
glimpse(df_durham_2015)
```

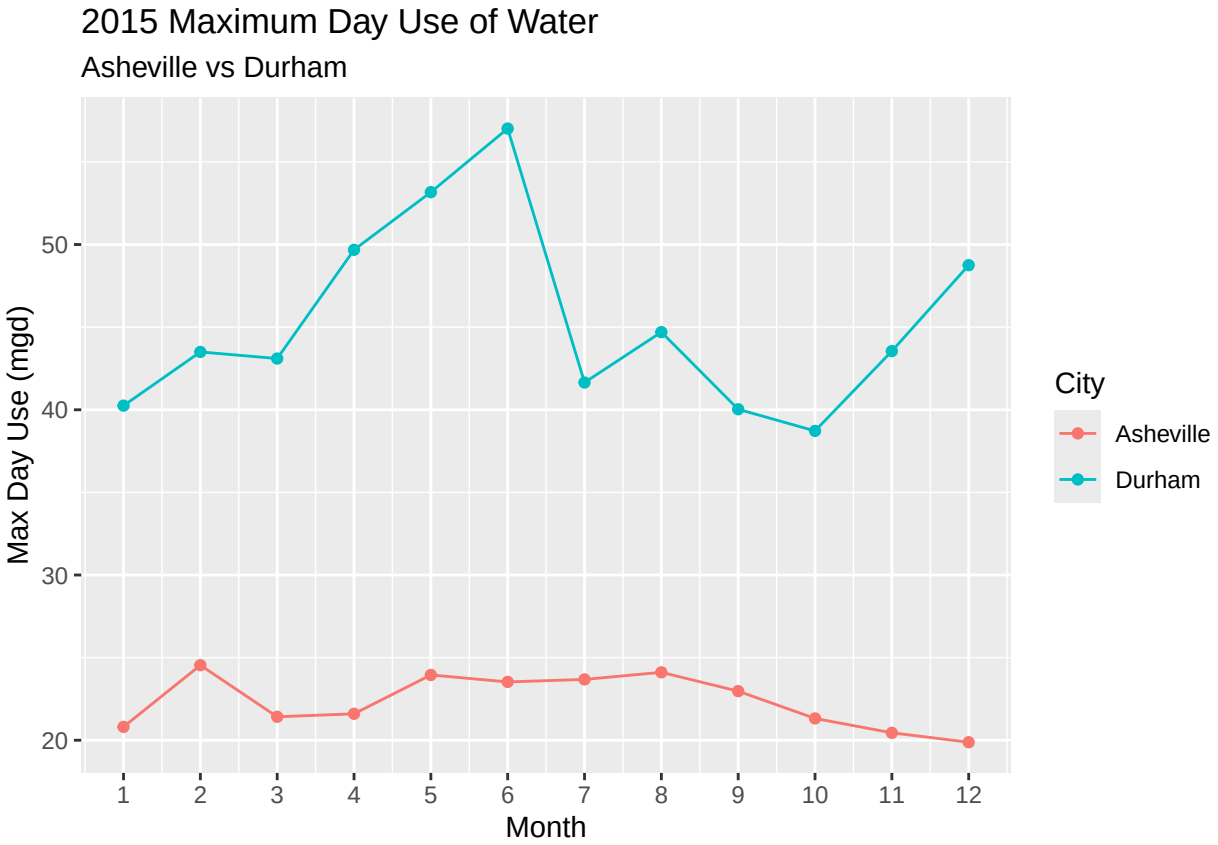
```
## Rows: 12
## Columns: 7
## $ Month               <dbl> 1, 5, 9, 2, 6, 10, 3, 7, 11, 4, 8, 12
## $ Year                <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 20~
## $ Maximum_Day_Use_mgd <dbl> 40.25, 53.17, 40.03, 43.50, 57.02, 38.72, 43.10, 4~
## $ Water_system_name   <chr> "Durham", "Durham", "Durham", "Durham", "Durham", ~
## $ PSWid               <chr> "03-32-010", "03-32-010", "03-32-010", "03-32-010"~
## $ Ownership           <chr> "Municipality", "Municipality", "Municipality", "M~
## $ Date                <date> 2015-01-01, 2015-05-01, 2015-09-01, 2015-02-01, 20~
```

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
df_asheville_2015 <- scrape_lwsp("01-11-010", 2015)
# checking result
glimpse(df_asheville_2015)
```

```
## Rows: 12
## Columns: 7
## $ Month               <dbl> 1, 5, 9, 2, 6, 10, 3, 7, 11, 4, 8, 12
## $ Year                <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 20~
## $ Maximum_Day_Use_mgd <dbl> 20.81, 23.95, 22.97, 24.54, 23.53, 21.32, 21.42, 2~
## $ Water_system_name   <chr> "Asheville", "Asheville", "Asheville", "Asheville"~
## $ PSWid               <chr> "01-11-010", "01-11-010", "01-11-010", "01-11-010"~
## $ Ownership           <chr> "Municipality", "Municipality", "Municipality", "M~
## $ Date                <date> 2015-01-01, 2015-05-01, 2015-09-01, 2015-02-01, 20~
```

```r
# combining dataset
df_joined_2015 <- rbind(df_durham_2015, df_asheville_2015)

# comparing water withdrawals
ggplot(df_joined_2015,
       aes(x = Month, y = Maximum_Day_Use_mgd,
           color = Water_system_name,
           group = Water_system_name)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks = 1:12) +
  labs(
    title = "2015 Maximum Day Use of Water",
    subtitle = "Asheville vs Durham",
    x = "Month",
    y = "Max Day Use (mgd)",
    color = "City")
```



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.

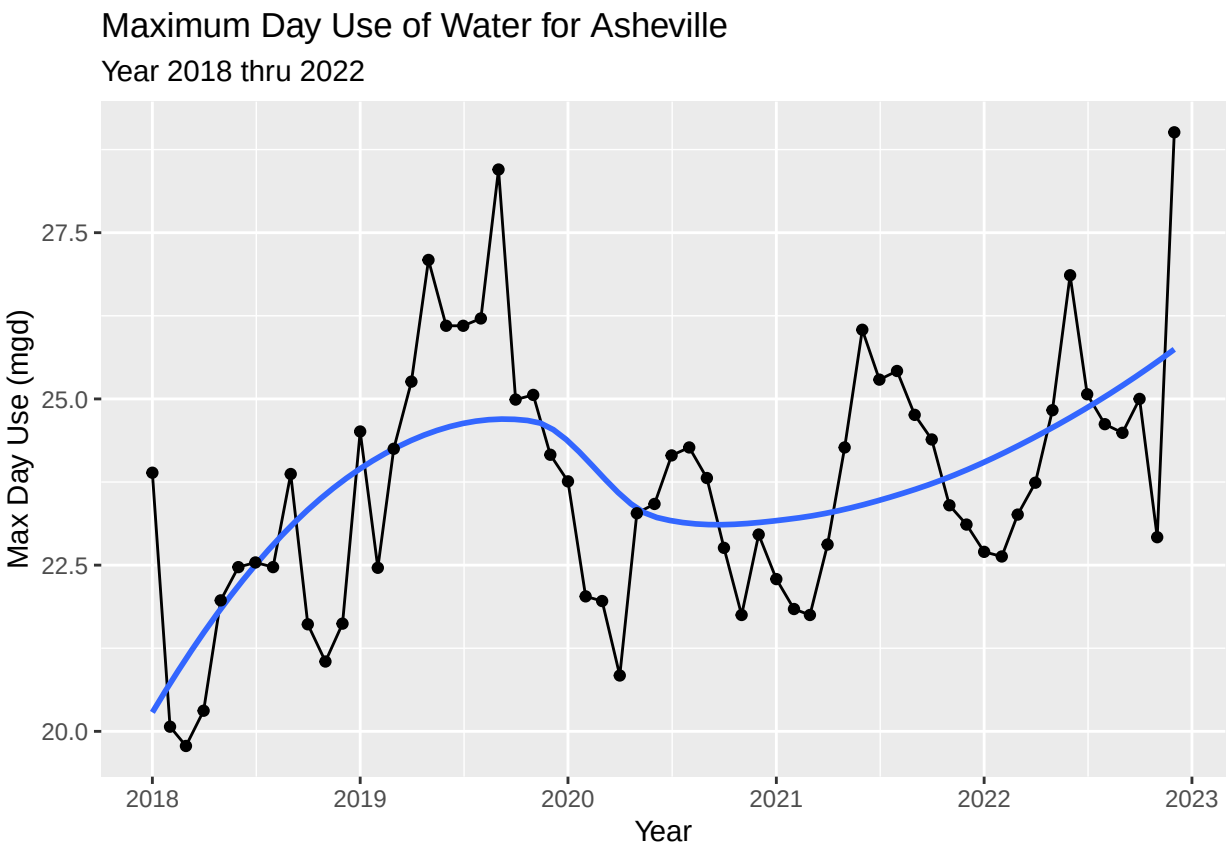
```
#9
years_of_interest = rep(2018:2022)
area_of_interest = '01-11-010'
dfs_desired <- lapply(X = years_of_interest,
                      FUN = scrape_lwsp,
                      pwsid = area_of_interest)

dfs_combined <- bind_rows(dfs_desired)

# checking results
glimpse(dfs_combined)
```

```
## Rows: 60
## Columns: 7
## $ Month               <dbl> 1, 5, 9, 2, 6, 10, 3, 7, 11, 4, 8, 12, 1, 5, 9, 2,~
## $ Year                <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 20~
## $ Maximum_Day_Use_mgd <dbl> 23.89, 21.97, 23.87, 20.07, 22.47, 21.61, 19.78, 2~
## $ Water_system_name   <chr> "Asheville", "Asheville", "Asheville", "Asheville"~
```

```
## $ PSWid                <chr> "01-11-010", "01-11-010", "01-11-010", "01-11-010"~
## $ Ownership            <chr> "Municipality", "Municipality", "Municipality", "M~
## $ Date                 <date> 2018-01-01, 2018-05-01, 2018-09-01, 2018-02-01, 2~
```

```
# creating a plot
ggplot(dfs_combined,
       aes(x = Date, y = Maximum_Day_Use_mgd)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  geom_point() +
  scale_x_date(
    date_breaks = "1 year",
    date_labels = "%Y"
  ) +
  labs(
    title = "Maximum Day Use of Water for Asheville",
    subtitle = "Year 2018 thru 2022",
    x = "Year",
    y = "Max Day Use (mgd)")
```

```
## `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? > Answer: > It seems like there is a trend of increasing use of water overtime. The anomaly year is 2019, when the water usage that year is exceptionally high. If we are ignoring the monthly variation, we can observe the increased usage of water for the same month from year 2018 to 2022.