Assignment 10: Data Scraping

Ardath Dixon

Total points:

OVERVIEW

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

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Fay_10_Data_Scraping.Rmd") prior to submission.

The completed exercise is due on Tuesday, April 6 at 11:59 pm.

Set up

1. Set up your session:

- Check your working directory
- Load the packages `tidyverse`, `rvest`, and any others you end up using.
- Set your ggplot theme

```
#1
getwd() ## setwd() to change if needed
```

```
## [1] "C:/Users/ardat/OneDrive/Documents/DataAnalytics/Environmental_Data_Analytics_2021"
```

```
library(tidyverse)
library(rvest)
```

```
## Warning: package 'rvest' was built under R version 4.0.4
```

```
library(lubridate)
library(dplyr)

mytheme <- theme_light(base_size = 12)+
  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 2019 Municipal Local Water Supply Plan (LWSP):

- Navigate to https://www.ncwater.org/WUDC/app/LWSP/search.php
- Change the date from 2020 to 2019 in the upper right corner.
- 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=2019

Indicate this website as the as the URL to be scraped.

```
#2
LWSP_Drhm2019_url <-
  "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2019"
LWSP_Drhm2019_website <- read_html(LWSP_Drhm2019_url)
```

3. The data we want to collect are listed below:

- From the "System Information" section:
- Water system name
- PSWID
- Ownership
- From the "Water Supply Sources" section:
- Maximum monthly withdrawals (MGD)

In the code chunk below scrape these values into the supplied variable names.

```
#3 Scrape data values for Water System Name, PWSID, & Ownership

WaterSystem_Drhm <- LWSP_Drhm2019_website %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>% html_text()

PWSID_Drhm <- LWSP_Drhm2019_website %>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)") %>% html_text()

Ownership_Drhm <- LWSP_Drhm2019_website %>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>% html_text()

MaxDayUse_monthly_Drhm <- LWSP_Drhm2019_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.)

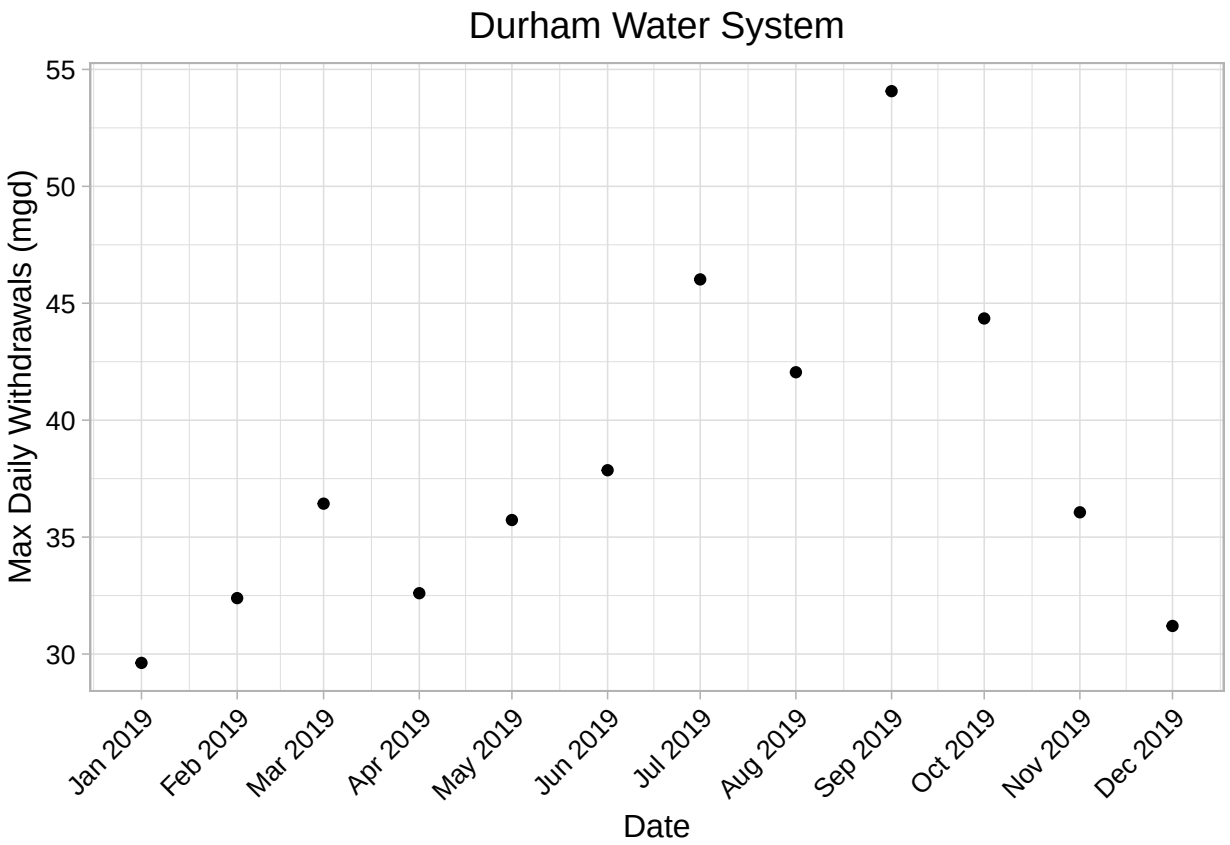
   NOTE: It's likely you won't be able to scrape the monthly widthrawal data in order. You can overcome this by creating a month column in the same order the data are scraped: Jan, May, Sept, Feb, etc. . .

5. Plot the max daily withdrawals across the months for 2019.

```
#4 create dataframe from scraped data
df_LWSP_Durham19 <- data.frame("Month" = c(1,5,9,2,6,10,3,7,11,4,8,12),
                          ## lists months in order of their display
                          "Year" = rep(2019, 12),
                          "WaterSystemName" = WaterSystem_Drhm,
                          "PWSID" = PWSID_Drhm,
                          "Ownership" = Ownership_Drhm,
                          "MaxDayUse_mgd" =
                            as.numeric(MaxDayUse_monthly_Drhm)) %>%
  mutate(GraphDate = my(paste(Month,"-",Year)))
```

```r
## reorder chronologically
df_LWSP_Durham19 <- df_LWSP_Durham19[order(df_LWSP_Durham19$Month),]


#5 plot max daily withdrawals across 2019 months
ggplot(df_LWSP_Durham19) +
  geom_point(aes(x = GraphDate, y = MaxDayUse_mgd))+
  scale_x_date(date_breaks = "1 month", date_labels = "%b %Y")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = .5))+
  ggtitle("Durham Water System")+
  labs(x = "Date", y = "Max Daily Withdrawals (mgd)" )
```



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 data scraped.

```r
#6.
scrape.it <- function(the_PWISD, the_year){
  NCwater_url <-
    paste0("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=",the_PWISD,
           "&year=",the_year)
  NCwater_website <- read_html(NCwater_url)

  WaterSys_tag <- "div+ table tr:nth-child(1) td:nth-child(2)"
  Ownership_tag <- "div+ table tr:nth-child(2) td:nth-child(4)"
  MaxDayUse_tag <-"th~ td+ td"
```

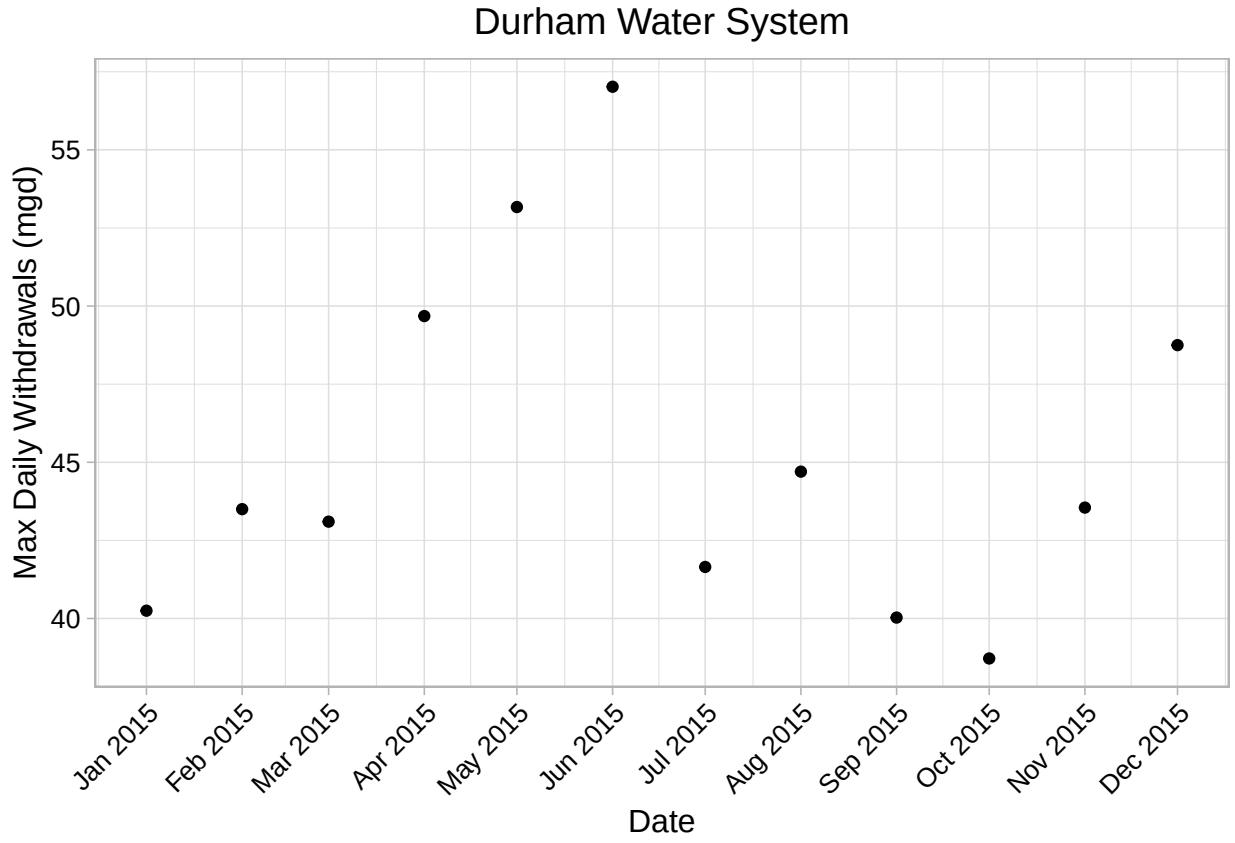
```
  water_system <- NCwater_website %>% html_nodes(WaterSys_tag) %>% html_text()
  the_ownership <- NCwater_website %>% html_nodes(Ownership_tag) %>% html_text()
  max_day_use <- NCwater_website %>% html_nodes(MaxDayUse_tag) %>% html_text()

  df_new <- data.frame("Month" = c(1,5,9,2,6,10,3,7,11,4,8,12),
                       "Year" = rep(the_year, 12),
                       "MaxDayUse_mgd" = as.numeric(max_day_use)) %>%
    mutate(WaterSystemName = !!water_system,
           Ownership = !!the_ownership,
           GraphDate = my(paste(Month,"-",Year)))
  df_new <- df_new[order(df_new$Month),] ## reordered chronologically
  return(df_new)
}
```

7. Use the function above to extract and plot max daily withdrawals for Durham for each month in 2015

```
#7
LWSP_df_1 <- scrape.it("03-32-010",2015) ## Durham's PWSID is 03-32-010

ggplot(LWSP_df_1) +
  geom_point(aes(x = GraphDate, y = MaxDayUse_mgd))+
  scale_x_date(date_breaks = "1 month", date_labels = "%b %Y")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  labs(x = "Date", y = "Max Daily Withdrawals (mgd)" )+
  ggtitle(paste0(LWSP_df_1$WaterSystemName," Water System"))+
  theme(plot.title = element_text(hjust = .5))
```

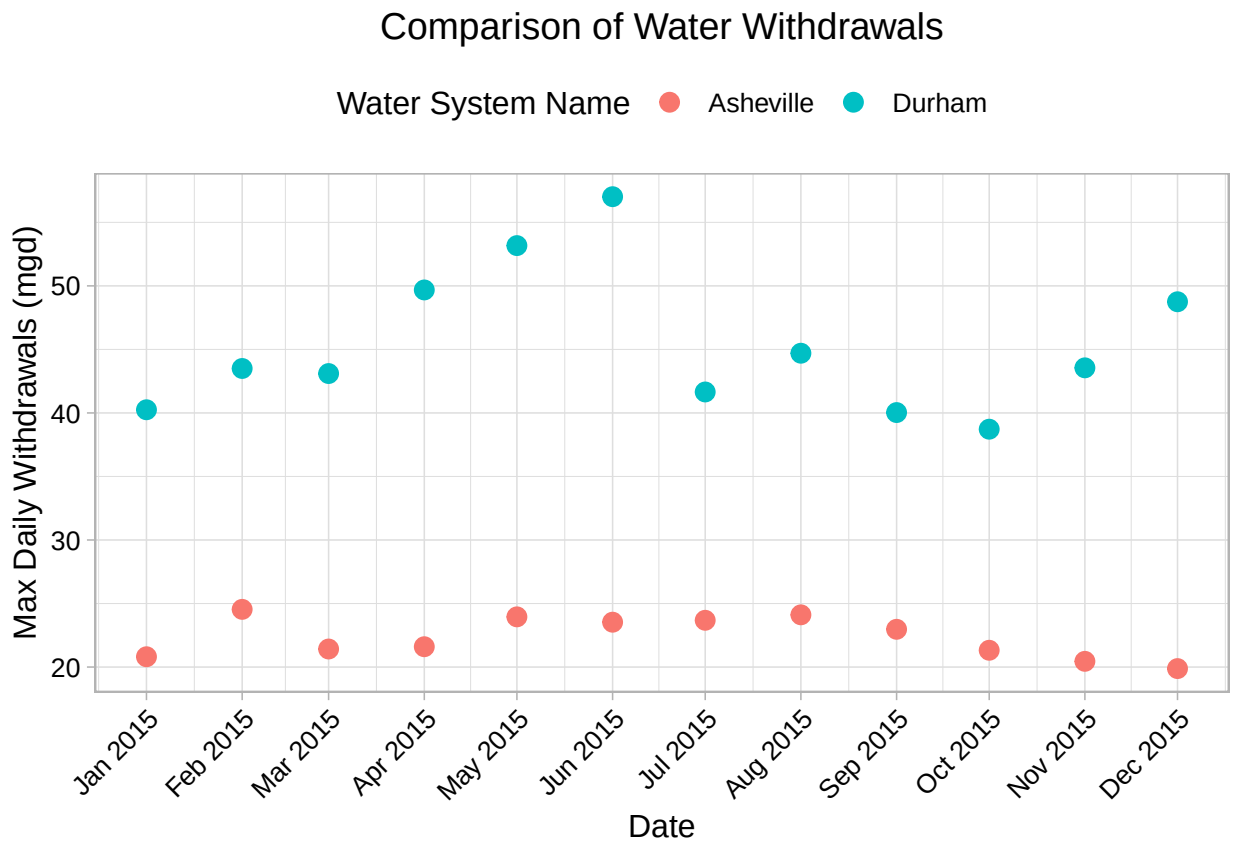


8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares the Asheville to Durham's water withdrawals.

```
#8
LWSP_df_2 <- scrape.it("01-11-010",2015) ## extract Asheville 2015 data
Combo_df_1_2 <- full_join(LWSP_df_1,LWSP_df_2)
```

```
## Joining, by = c("Month", "Year", "MaxDayUse_mgd", "WaterSystemName", "Ownership", "GraphDate")
```

```
ComboGraph <- ggplot(Combo_df_1_2) +
  geom_point(aes(x = GraphDate, y = MaxDayUse_mgd,
                 color = WaterSystemName), size = 3)+
  scale_x_date(date_breaks = "1 month", date_labels = "%b %Y")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  labs(x = "Date", y = "Max Daily Withdrawals (mgd)",
       color = "Water System Name")+
  ggtitle("Comparison of Water Withdrawals")+
  theme(plot.title = element_text(hjust = .5))
ComboGraph
```



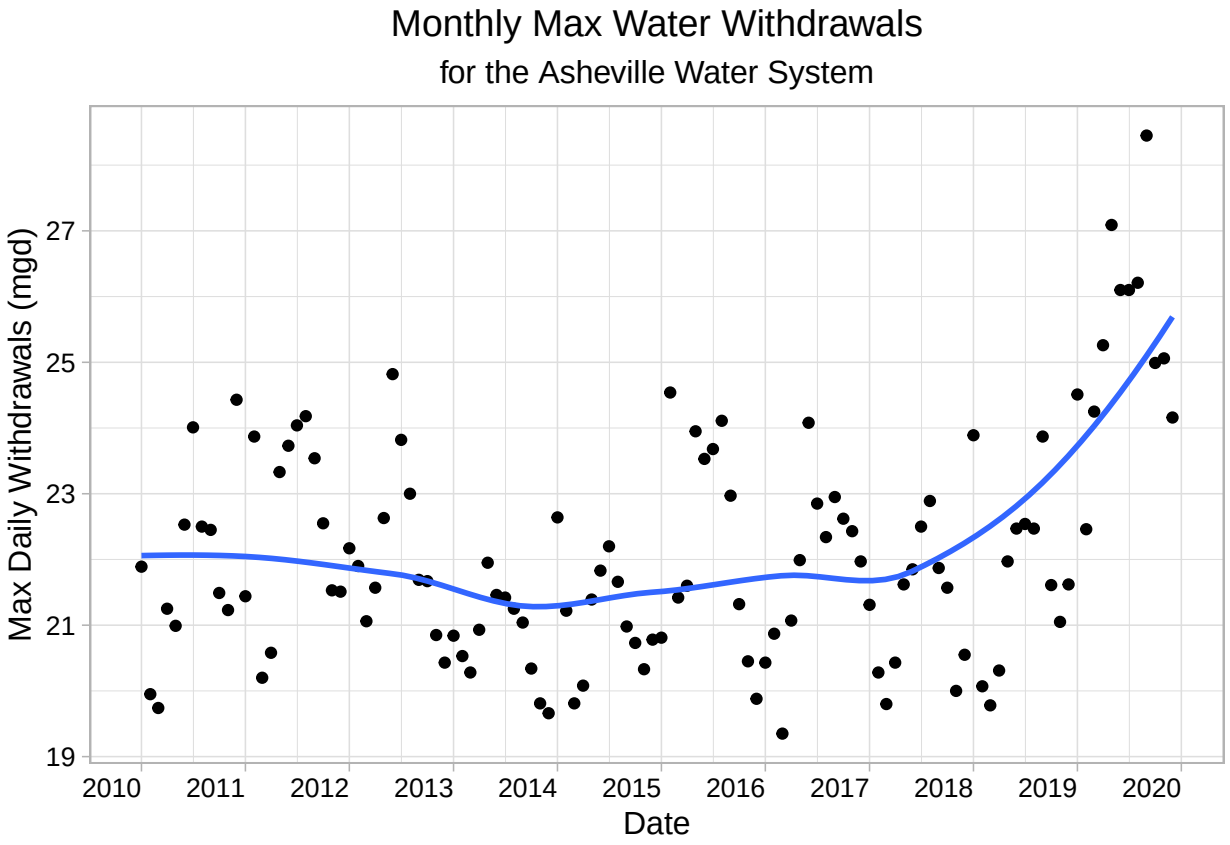
9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

```
#9
the_years = rep(2010:2019)
PWISD_number = "01-11-010"
```

```
the_dfs <- map(the_years, scrape.it, the_PWISD = PWISD_number)
the_df <- bind_rows(the_dfs)

AshevilleGraph <- ggplot(the_df) +
  geom_point(aes(x = GraphDate, y = MaxDayUse_mgd))+
  geom_smooth(aes(x = GraphDate, y = MaxDayUse_mgd), se=FALSE)+
  scale_x_date(date_breaks = "1 year", date_labels = "%Y")+
  theme(axis.text.x = element_text(hjust = 1))+
  labs(x = "Date", y = "Max Daily Withdrawals (mgd)")+
  ggtitle("Monthly Max Water Withdrawals","for the Asheville Water System")+
  theme(plot.title = element_text(hjust = .5),
        plot.subtitle = element_text(hjust = 0.5))
AshevilleGraph
```

```
## `geom_smooth()` using method = 'loess' and 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?

In looking at the plot, Asheville appears to have a trend of increasing water usage over time. The trend line increases drastically from 2018 to 2020, showing an overall rise in maximum daily water withdrawals.