Assignment 09: Data Scraping

Ben Egan

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_09_Data_Scraping.Rmd") prior to submission.

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()
```

```
## [1] "C:/Users/benja/OneDrive/Documents/R/win-library/Environmental_Data_Analytics_2022/Assignments"
```

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

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 2021 to 2020 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=2020

Indicate this website as the as the URL to be scraped. (In other words, read the contents into an `rvest` webpage object.)

```
#2
theURL <- 'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2020'

webpage <- read_html(theURL)
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
- PSWID
- Ownership
- From the "3. Water Supply Sources" section:
- Max Daily Use (MDU) - for each month

In the code chunk below scrape these values, assigning them to three 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, with the first value being 36.0100.

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

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

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

max.withdrawals.mgd <- webpage %>%
  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 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 2020

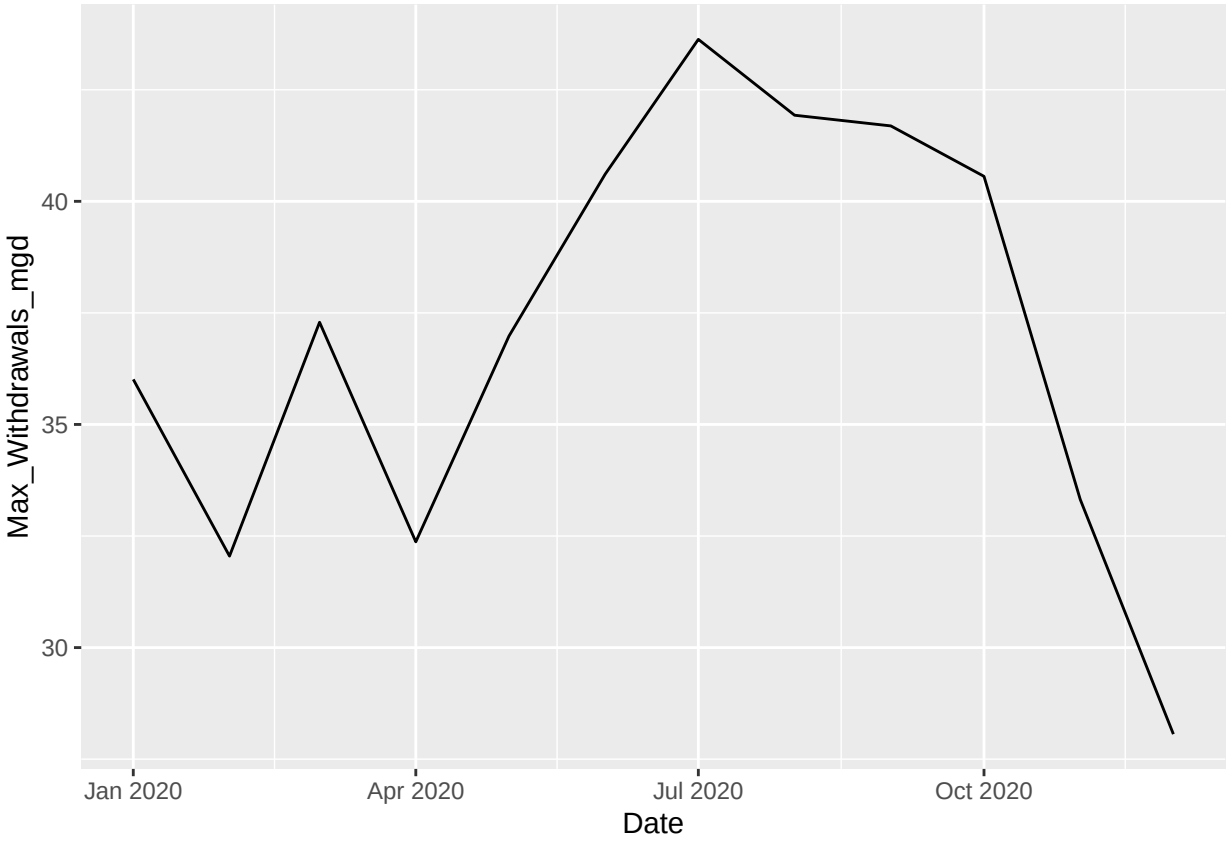
```
#4
df_withdrawals <- data.frame("Month" = c("1","5","9","2",
                                                    "6","10","3","7",
                                                    "11", "4", "8","12" ),
                             "Year" = rep("2020"),
                             "PSWID" = rep(pswid),
                             "Ownership" = rep(ownership),
                             "Max_Withdrawals_mgd" =
                               as.numeric(max.withdrawals.mgd),
                             "Water_System_Name" = rep(water.system.name))

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

#5

df_withdrawals_plot <-ggplot(df_withdrawals, aes(x=Date, y=Max_Withdrawals_mgd)) +
  geom_line()

print(df_withdrawals_plot)
```



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 scraped**.

```
#6.
Base_URL <- 'https://www.ncwater.org/WUDC/app/LWSP/report.php?'
Base_pwisd <- 'pwsid=03-32-010'
Base_year <- 2015

scrape.it <- function(Base_pwisd,Base_year){

the_scrape_url <- paste0(Base_URL, Base_pwisd,'&year=', Base_year)

print(the_scrape_url)

the_website <- read_html(the_scrape_url)

water.system.name_f <- the_website %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()

pwsid_f <- the_website %>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
  html_text()
```

```
ownership_f <- the_website %>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%
  html_text()

max.withdrawals.mgd_f <- the_website %>% html_nodes("th~ td+ td") %>%
  html_text()

df_withdrawals_f <- data.frame("Month" = c("1","5","9","2",
                                           "6","10","3","7",
                                           "11", "4", "8","12" ),
                               "Year" = rep(Base_year),
                               "PWSID" = rep(pwsid_f),
                               "Ownership" = rep(ownership_f),
                               "Max_Withdrawals_mgd" =
                                 as.numeric(max.withdrawals.mgd_f),
                               "Water_System_Name" = rep(water.system.name_f))

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

Sys.sleep(1)

return(df_withdrawals_f)
}
```

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

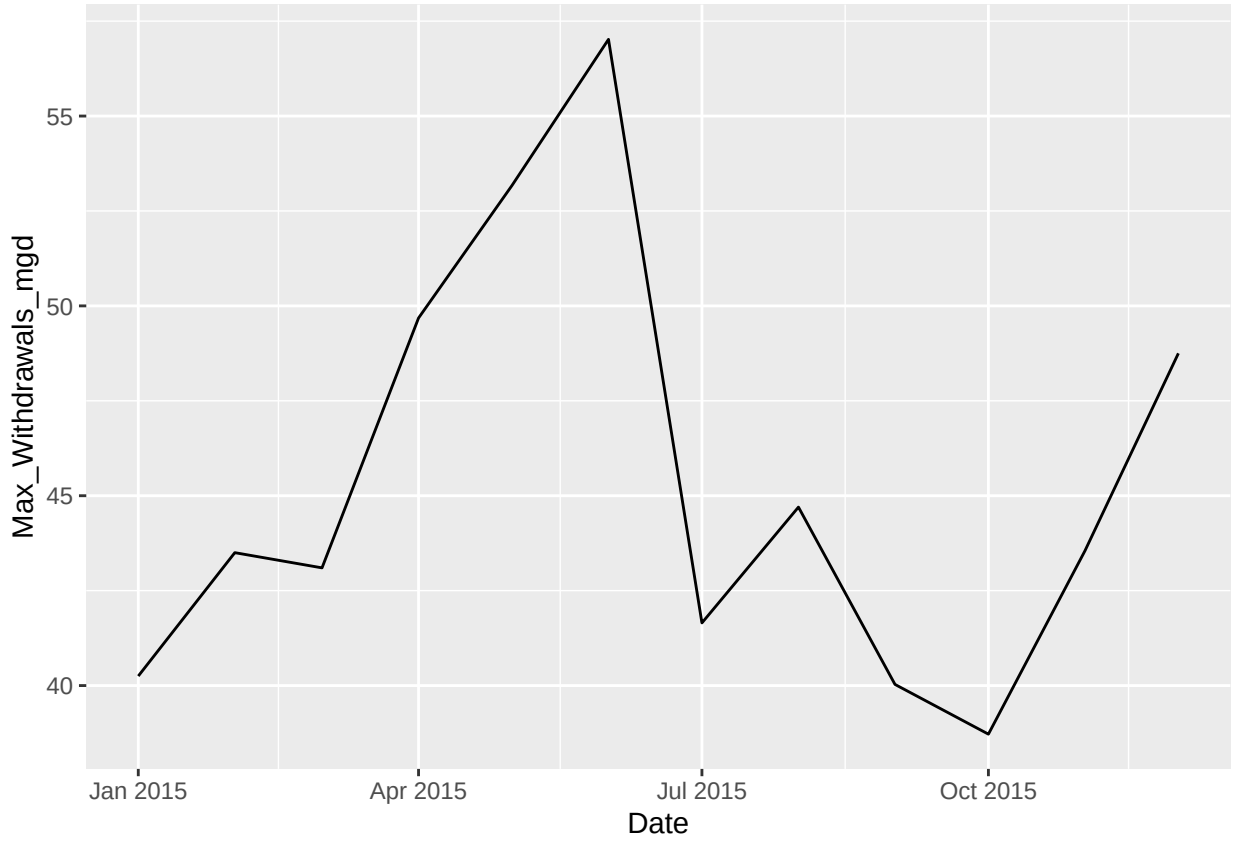
```
#7

df_2015 <- scrape.it('pwsid=03-32-010',2015)
```

```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2015"
```

```
df_2015_plot <-ggplot(df_2015, aes(x=Date, y=Max_Withdrawals_mgd)) +
  geom_line()

print(df_2015_plot)
```



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

df_2015_ash <- scrape.it('pwsid=01-11-010',2015)
```

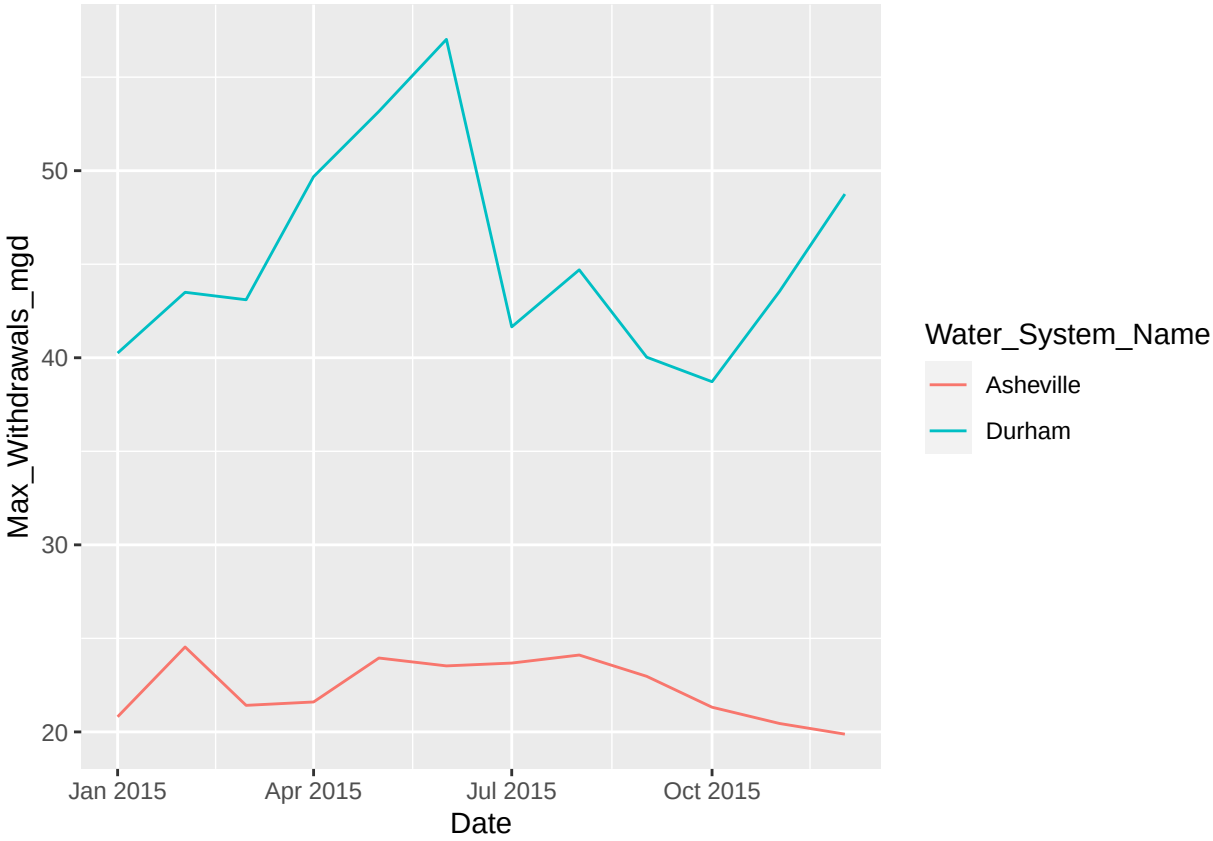
```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2015"
```

```
view(df_2015_ash)

df_withdrawals_combined <- rbind(df_2015,df_2015_ash)

df_withdrawals_plot_Combined <-ggplot(df_withdrawals_combined,
                                      aes(x=Date, y=Max_Withdrawals_mgd,
                                          color = Water_System_Name)) +
  geom_line()

print(df_withdrawals_plot_Combined)
```



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)
the_pwsid = 'pwsid=01-11-010'

the_dfs_ash <- lapply(X= the_years, FUN = scrape.it, Base_pwisd=the_pwsid)
```

```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2010"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2011"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2012"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2013"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2014"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2015"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2016"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2017"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2018"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2019"
```

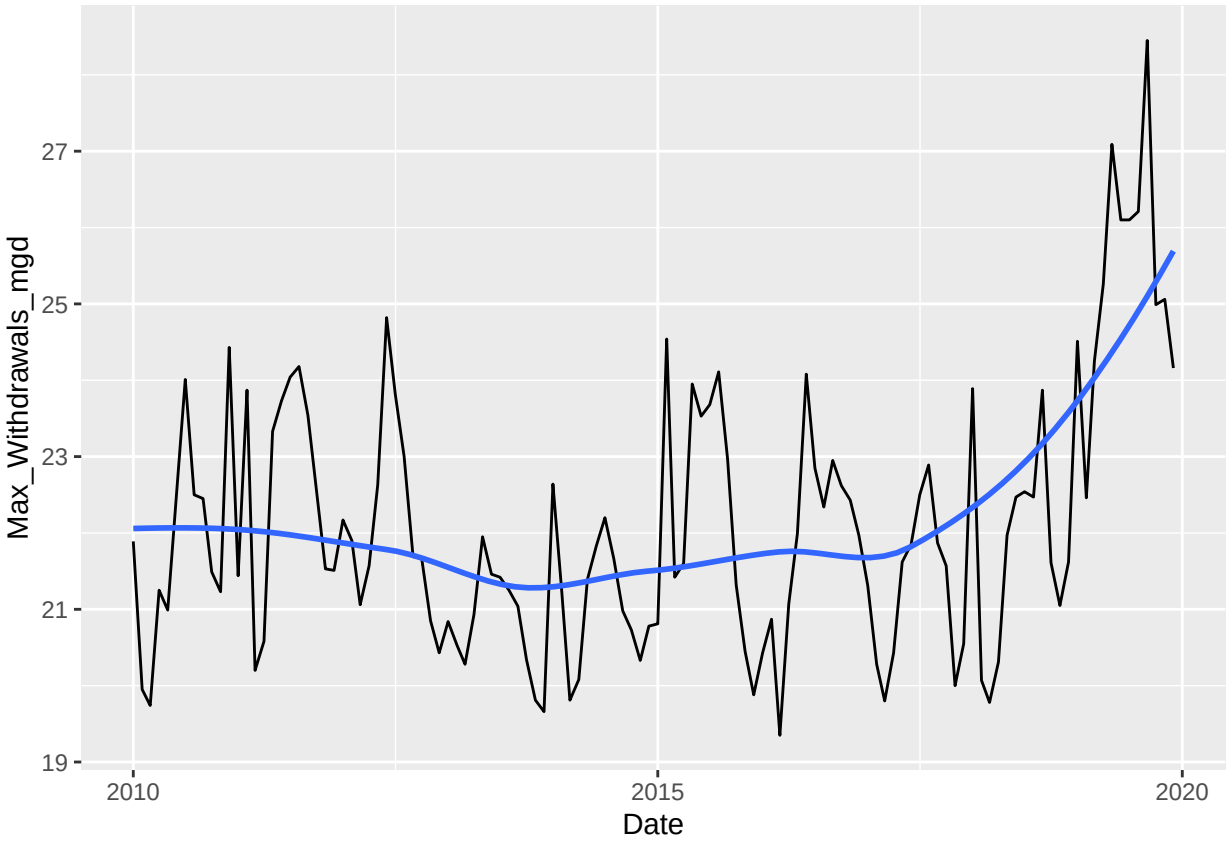
```
the_df_combined <- bind_rows(the_dfs_ash)

the_df_combined_plot <-ggplot(the_df_combined,
                                    aes(x=Date, y=Max_Withdrawals_mgd,
                                        )) +
```

```
  geom_line() + geom_smooth(method="loess",se=FALSE)

print(the_df_combined_plot)
```

```
## `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?

from 2010 to 2015 the water usage is stagnant. However, from 2015 to 2020, water usage is steadily increasing year over year.