Assignment 09: Data Scraping

Yikai Jing

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] "/Users/me/Environmental_Data_Analytics_2022/Assignments"
```

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

mytheme <- theme_classic(base_size = 12) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
```

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=2020

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=2020')
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:
- Average Daily Use (MGD) - 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()
water.system.name
```

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

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

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

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

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

```
max.withdrawals.mgd <- webpage %>%
  html_nodes("th~ td+ td") %>%
  html_text()
max.withdrawals.mgd
```

```
##  [1] "36.0100" "36.9800" "41.6900" "32.0500" "40.6100" "40.5600" "37.2900"
##  [8] "43.6300" "33.3200" "32.3700" "41.9300" "28.0600"
```

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.

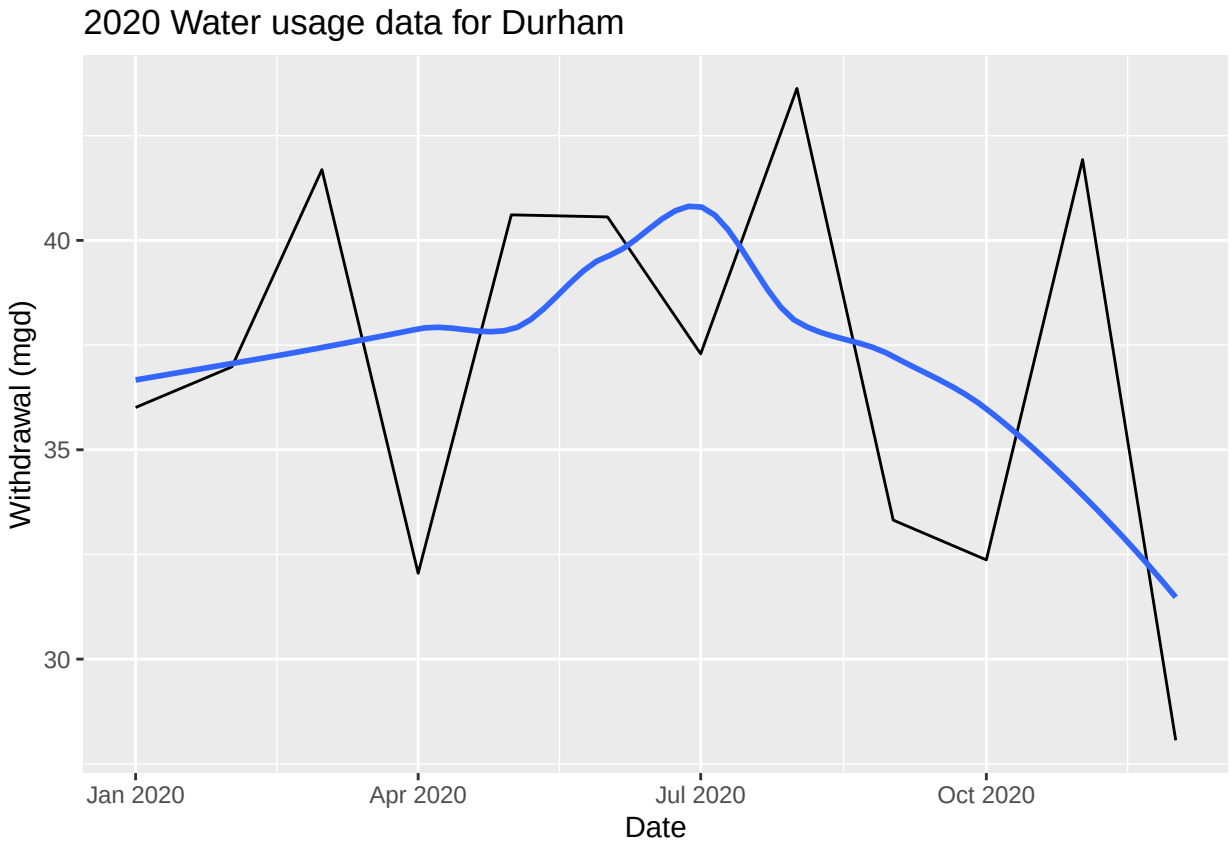
NOTE: It's likely you won't be able to scrape the monthly widthrawal data in order. You can overcome this by creating a month column in the same order the data are scraped: Jan, May, Sept, Feb, etc...

5. Plot the max daily withdrawals across the months for 2020

```
#4
df_withdrawals <- data.frame("Month" = rep(1:12),
                             "Year" = rep(2020,12),
                             "Max_Withdrawals_mgd" = as.numeric(max.withdrawals.mgd))
df_withdrawals <- df_withdrawals %>%
  mutate(Water_System_name = !!water.system.name,
         Ownership = !!ownership,
         Date = my(paste(Month,"-",Year)))


#5
ggplot(df_withdrawals,aes(x=Date,y=Max_Withdrawals_mgd)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2020 Water usage data for",water.system.name),
       y="Withdrawal (mgd)",
       x="Date")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



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.
scrape.it <- function(the_pswid, the_year){
  the_website <- read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',
                                  the_pswid, '&year=', the_year))

  water_system_name_tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'
  ownership_tag <- 'div+ table tr:nth-child(2) td:nth-child(4)'
  the_data_tag <- 'th~ td+ td'

  water_system_name <- the_website %>% html_nodes(water_system_name_tag) %>% html_text()
  ownership <- the_website %>%   html_nodes(ownership_tag) %>%  html_text()
  max_withdrawals <- the_website %>% html_nodes(the_data_tag) %>% html_text()

  df_withdrawals <- data.frame("Month" = rep(1:12),
                               "Year" = rep(the_year,12),
                               "Max_Withdrawals_mgd" = as.numeric(max_withdrawals)) %>%
    mutate(Water_System_name = !!water_system_name,
         Ownership = !!ownership,
         Date = my(paste(Month,"-",Year)))

  print(paste('The Pswid =', the_pswid, ', The Year =', the_year))
  return(df_withdrawals)
}
```

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

```
#7
scrape.it('03-32-010', 2015)
```

```
## [1] "The Pswid = 03-32-010 , The Year = 2015"
```

```
##    Month Year Max_Withdrawals_mgd Water_System_name    Ownership       Date
## 1      1 2015               40.25           Durham Municipality 2015-01-01
## 2      2 2015               53.17           Durham Municipality 2015-02-01
## 3      3 2015               40.03           Durham Municipality 2015-03-01
## 4      4 2015               43.50           Durham Municipality 2015-04-01
## 5      5 2015               57.02           Durham Municipality 2015-05-01
## 6      6 2015               38.72           Durham Municipality 2015-06-01
## 7      7 2015               43.10           Durham Municipality 2015-07-01
## 8      8 2015               41.65           Durham Municipality 2015-08-01
## 9      9 2015               43.55           Durham Municipality 2015-09-01
## 10    10 2015               49.68           Durham Municipality 2015-10-01
## 11    11 2015               44.70           Durham Municipality 2015-11-01
## 12    12 2015               48.75           Durham Municipality 2015-12-01
```

```
the_df_Durham <- scrape.it('03-32-010', 2015)
```

```
## [1] "The Pswid = 03-32-010 , The Year = 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 the Asheville to Durham’s water withdrawals.

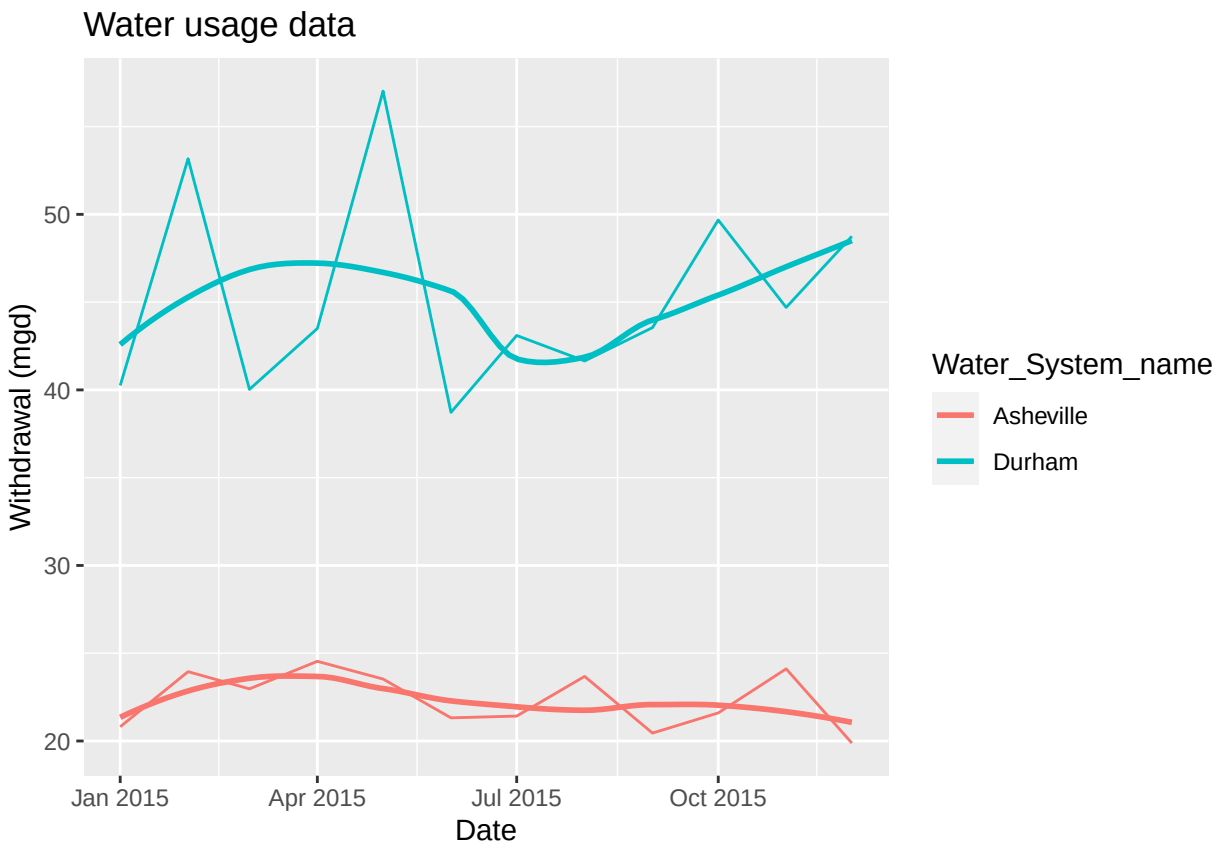
```
#8
the_df_Asheville <- scrape.it('01-11-010', 2015)
```

```
## [1] "The Pswid = 01-11-010 , The Year = 2015"
```

```
the_df <- bind_rows(the_df_Durham, the_df_Asheville)

ggplot(the_df,aes(x=Date,y=Max_Withdrawals_mgd,color=Water_System_name)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = "Water usage data",
       y="Withdrawal (mgd)",
       x="Date")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



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(2010:2020)
my_pswid = '01-11-010'
the_dfs_Asheville <- map(the_years,scrape.it,the_pswid=my_pswid)
```

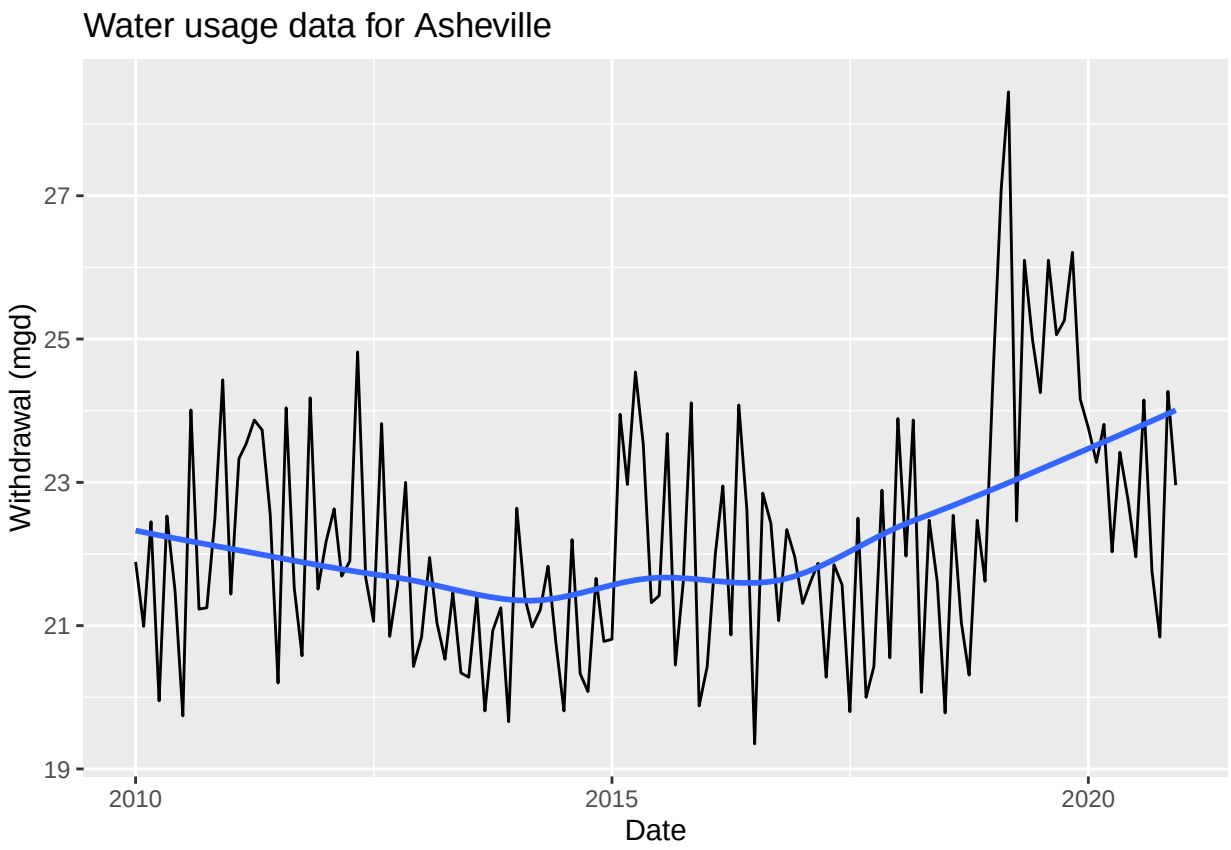
```
## [1] "The Pswid = 01-11-010 , The Year = 2010"
## [1] "The Pswid = 01-11-010 , The Year = 2011"
## [1] "The Pswid = 01-11-010 , The Year = 2012"
## [1] "The Pswid = 01-11-010 , The Year = 2013"
## [1] "The Pswid = 01-11-010 , The Year = 2014"
```

```
## [1] "The Pswid = 01-11-010 , The Year = 2015"
## [1] "The Pswid = 01-11-010 , The Year = 2016"
## [1] "The Pswid = 01-11-010 , The Year = 2017"
## [1] "The Pswid = 01-11-010 , The Year = 2018"
## [1] "The Pswid = 01-11-010 , The Year = 2019"
## [1] "The Pswid = 01-11-010 , The Year = 2020"
```

```
the_dfs_Asheville <- bind_rows(the_dfs_Asheville)


ggplot(the_dfs_Asheville, aes(x=Date,y=Max_Withdrawals_mgd)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("Water usage data for Asheville"),
       y="Withdrawal (mgd)",
       x="Date")
```

```
## `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? Since 2015, the max water usage tends to increase.