Assignment 09: Data Scraping

Lehe, Xu

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] "D:/Documents/Environmental_Data_Analytics_2022"

library(tidyverse)
library(lubridate)

#install.packages("rvest")
library(rvest)

## Warning: 程辑包'rvest'是用 R 版本 4.1.3 来建造的

# Set theme
mytheme <- theme_classic() +
  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=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="t
op" 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()
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()
```

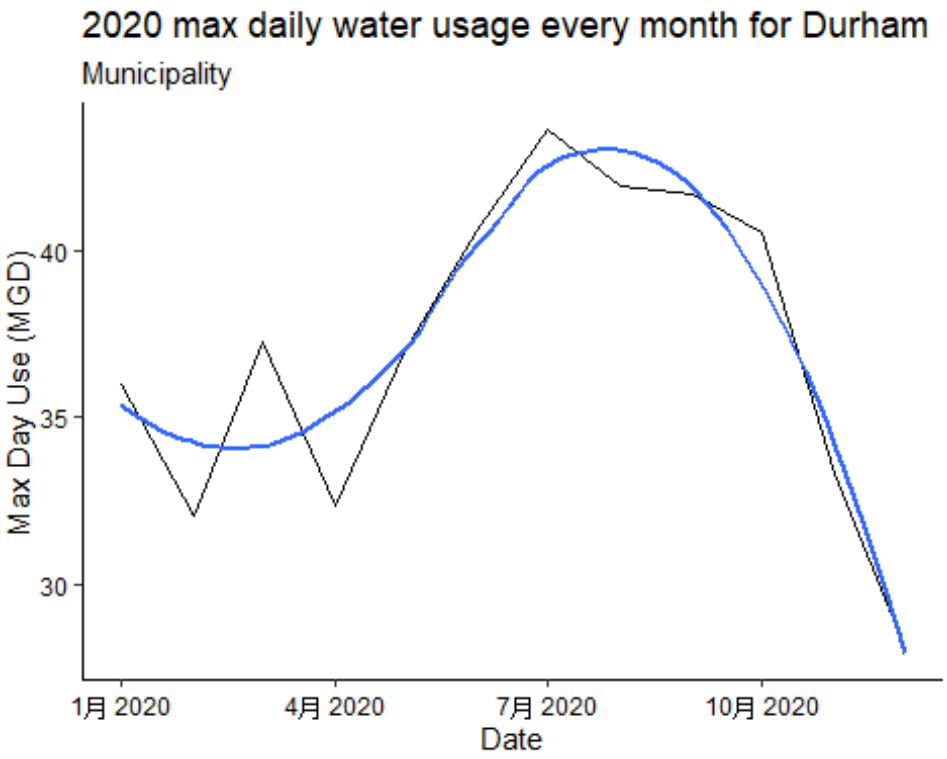
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   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
the_df <- data.frame(
    "water_system_name" = rep(water.system.name),
    "pswid"= rep(pswid),
    "ownership"=rep(ownership),
    "max.withdrawals.mgd"= as.numeric(max.withdrawals.mgd),
    "month"= c("1","5","9","2","6","10","3","7","11","4","8","12"),
    "year"=rep(2020))

the_df <- the_df %>%
  mutate(Date = my(paste(month,"-",year)))
#5
ggplot(the_df,aes(x=Date,y=max.withdrawals.mgd)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title ="2020 max daily water usage every month for Durham",
       subtitle = ownership,
       y="Max Day Use (MGD)",
       x="Date")

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



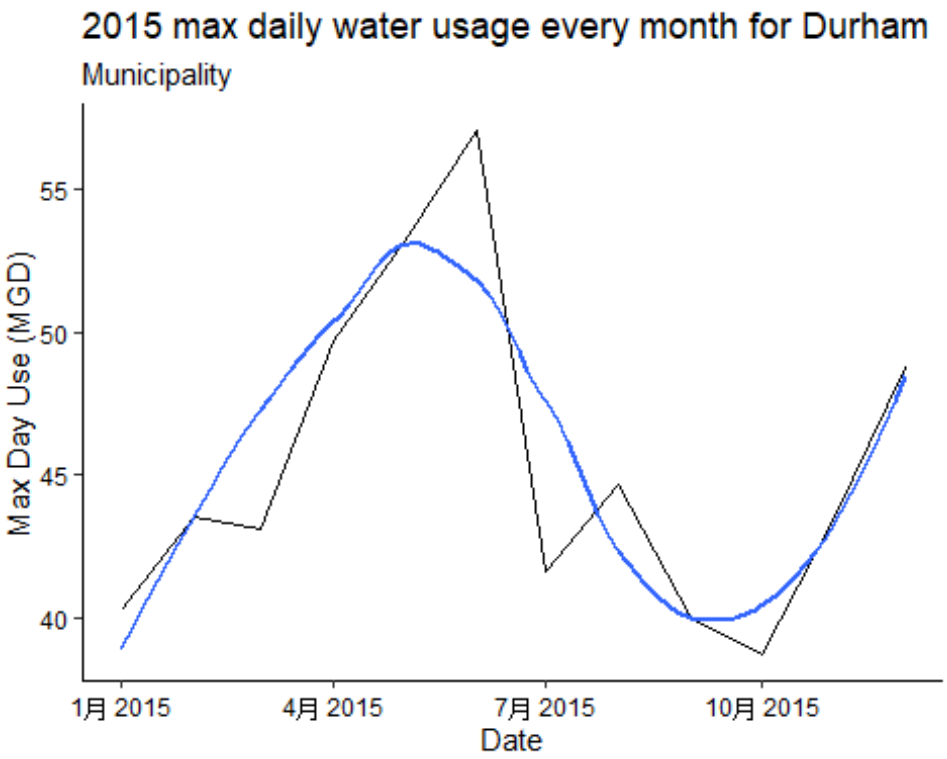
6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site scraped**.

```
#6.
scrape.it <- function(pwsid,the_year){

webpage <- read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/repo
rt.php?',
                            'pwsid=', pwsid, '&year=', the_year))

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()

the_df <- data.frame(
    "water_system_name" = rep(water.system.name),
```

```
    "pswid"= rep(pswid),
    "ownership"=rep(ownership),
    "max.withdrawals.mgd"= as.numeric(max.withdrawals.mgd),
    "month"= c("1","5","9","2","6","10","3","7","11","4","8","12"),
    "year"=rep(the_year)) %>%
  mutate(Date = my(paste(month,"-",year)))

return(the_df)
}
```

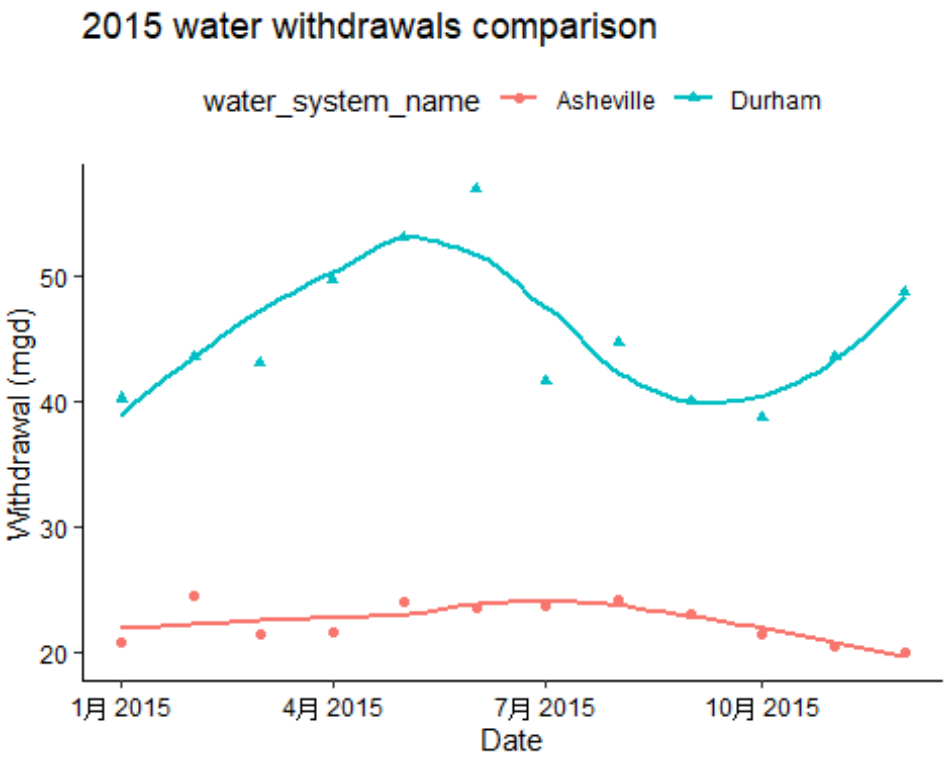
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID='03-32-010') for each month in 2015

```
#7
Durham_df_2015 <- scrape.it('03-32-010', 2015)
view(Durham_df_2015)

ggplot(Durham_df_2015,aes(x=Date,y=max.withdrawals.mgd)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = "2015 max daily water usage every month for Durham",
       subtitle = ownership,
       y="Max Day Use (MGD)",
       x="Date")

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



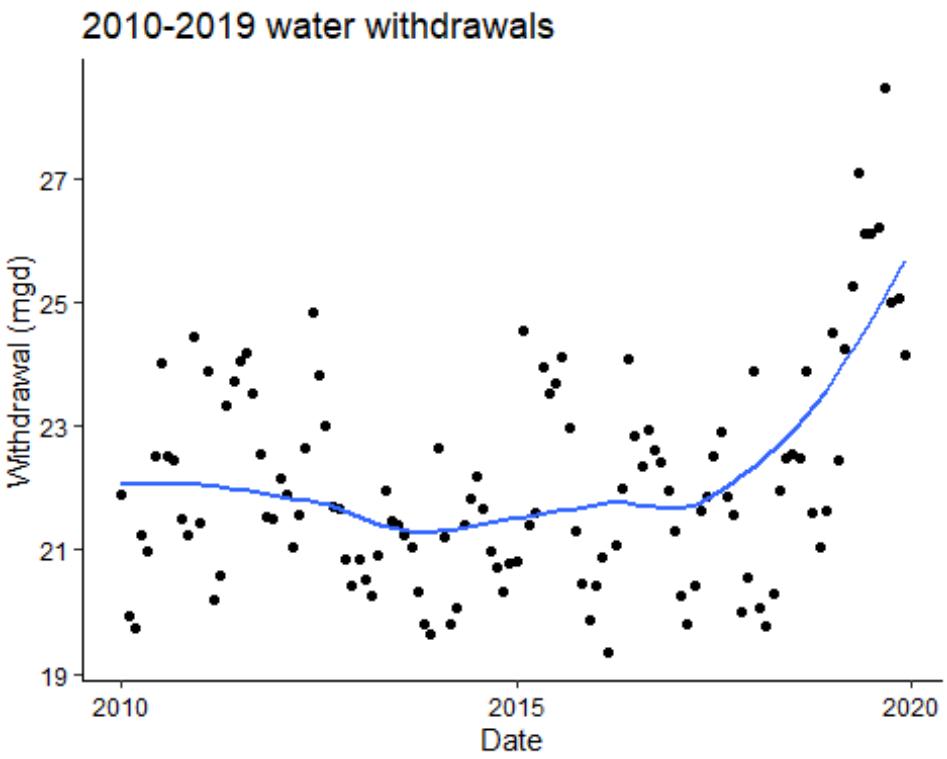
8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares the Asheville to Durham's water withdrawals.

```
#8
Asheville_df_2015 <- scrape.it('01-11-010', 2015)
view(Asheville_df_2015)

join_2015 <- full_join(Asheville_df_2015,Durham_df_2015)

## Joining, by = c("water_system_name", "pswid", "ownership",
## "max.withdrawals.mgd", "month", "year", "Date")

compare <-
  ggplot(join_2015, aes(x = Date, y = max.withdrawals.mgd, color = wate
r_system_name, shape = water_system_name)) +
  geom_point() +
  geom_smooth(se=FALSE)+
    labs(title = "2015 water withdrawals comparison",
       y="Withdrawal (mgd)",
       x="Date")
print(compare)

## `geom_smooth()` using method = 'loess' and 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 = c(2010:2019)

All_Asheville <- cross2('01-11-010', the_years) %>%
  map(lift(scrape.it)) %>%
  bind_rows()


Asheville <-
  ggplot(All_Asheville, aes(x = Date, y = max.withdrawals.mgd)) +
  geom_point() +
  geom_smooth(se=FALSE)+
    labs(title = "2010-2019 water withdrawals",
       y="Withdrawal (mgd)",
       x="Date")
print(Asheville)

## `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?

Yes, Asheville have a trend in water usage over time. From 2010-2015, the water usage was gradually decreasing, but after 2015 the water usage started increasing. Especially, afte 2018, water usage increased rapidly. Overall, water usage shows an increasing trend over time.