Assignment 09: Data Scraping

Zhiyuan Chen

OVERVIEW

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

Directions

1. Rename this file `<FirstLast>_A09_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 to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.

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:/Rfiles/EDA-Fall2022"
```

```
#install.packages("tidyverse")
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.2
```

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

```
## Warning: package 'rvest' was built under R version 4.2.2
```

```
#install.packages("lubridate")
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.2.2

## Warning: package 'timechange' was built under R version 4.2.2
```

```
mytheme <- theme_classic(base_size = 14) +
  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 2021 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=2021

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=2021')
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:
- Maximum Daily 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), with the first value being "27.6400".

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

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

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

```
## [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 = as.numeric(max.withdrawals.mgd)
max.withdrawals.mgd
```

```
##  [1] 27.64 41.79 36.72 27.97 37.95 42.24 30.54 43.62 31.28 33.76 46.08 29.78
```

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.

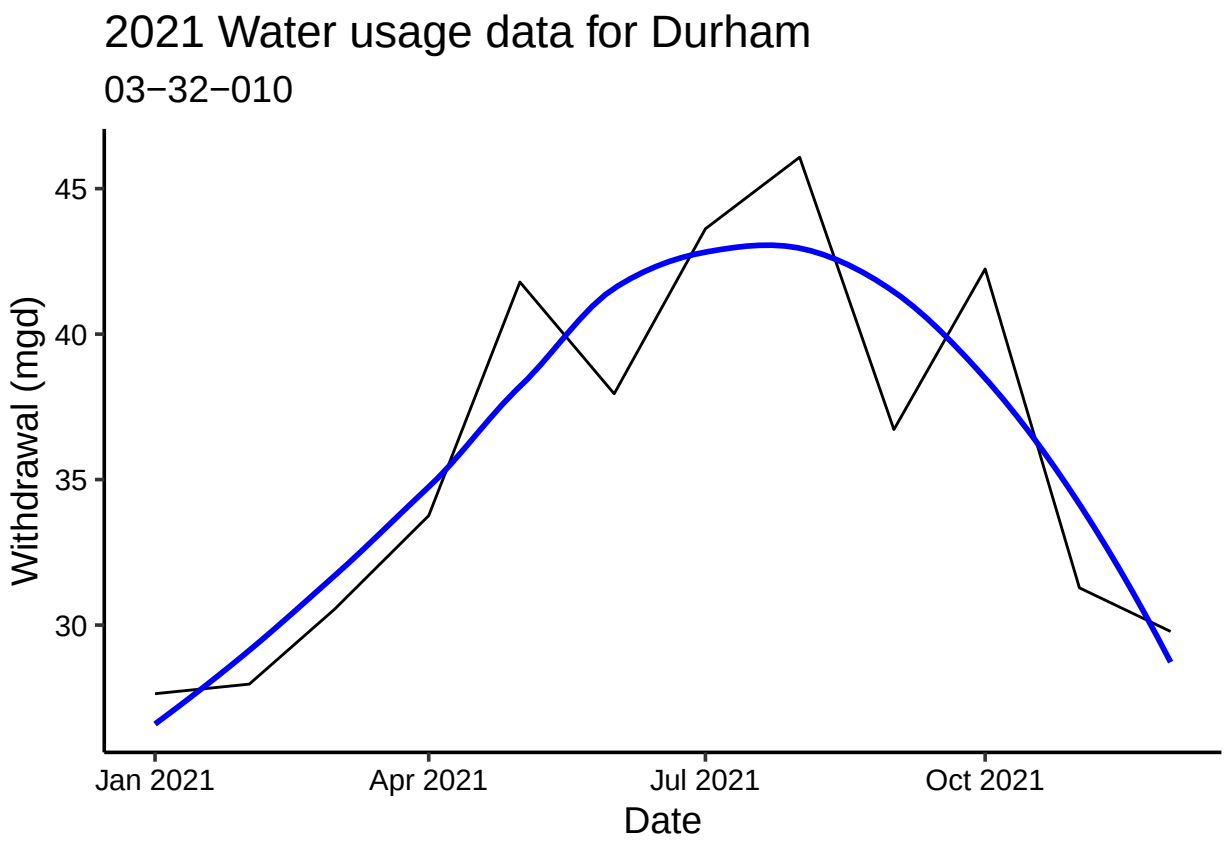
   NOTE: It's likely you won't be able to scrape the monthly widthrawal data in chronological order. You can overcome this by creating a month column manually assigning values in the order the data are scraped: "Jan", "May", "Sept", "Feb", etc...

5. Create a line plot of the maximum daily withdrawals across the months for 2021

```
#4
df_withdrawals <- data.frame("Month" = c(1,5,9,2,6,10,3,7,11,4,8,12),
                             "Year" = rep(2021,12),
                             "max.withdrawals.mgd" = as.numeric(max.withdrawals.mgd)) %>%
  mutate(water.system.name = !!water.system.name,
         pwsid = !!pwsid,
         ownership = !!ownership,
         Date = paste0(Month,"-",1,"-",Year))
df_withdrawals$Date <- as.Date(df_withdrawals$Date,format = "%m-%d-%Y")
#5
ggplot(df_withdrawals,aes(x=Date,y=max.withdrawals.mgd)) +
  geom_line(aes(group=1)) +
  geom_smooth(method="loess",se=FALSE, color="blue") +
```

```
  labs(title = paste("2021 Water usage data for",water.system.name),
       subtitle = pwsid,
       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 (pwsid) scraped**.

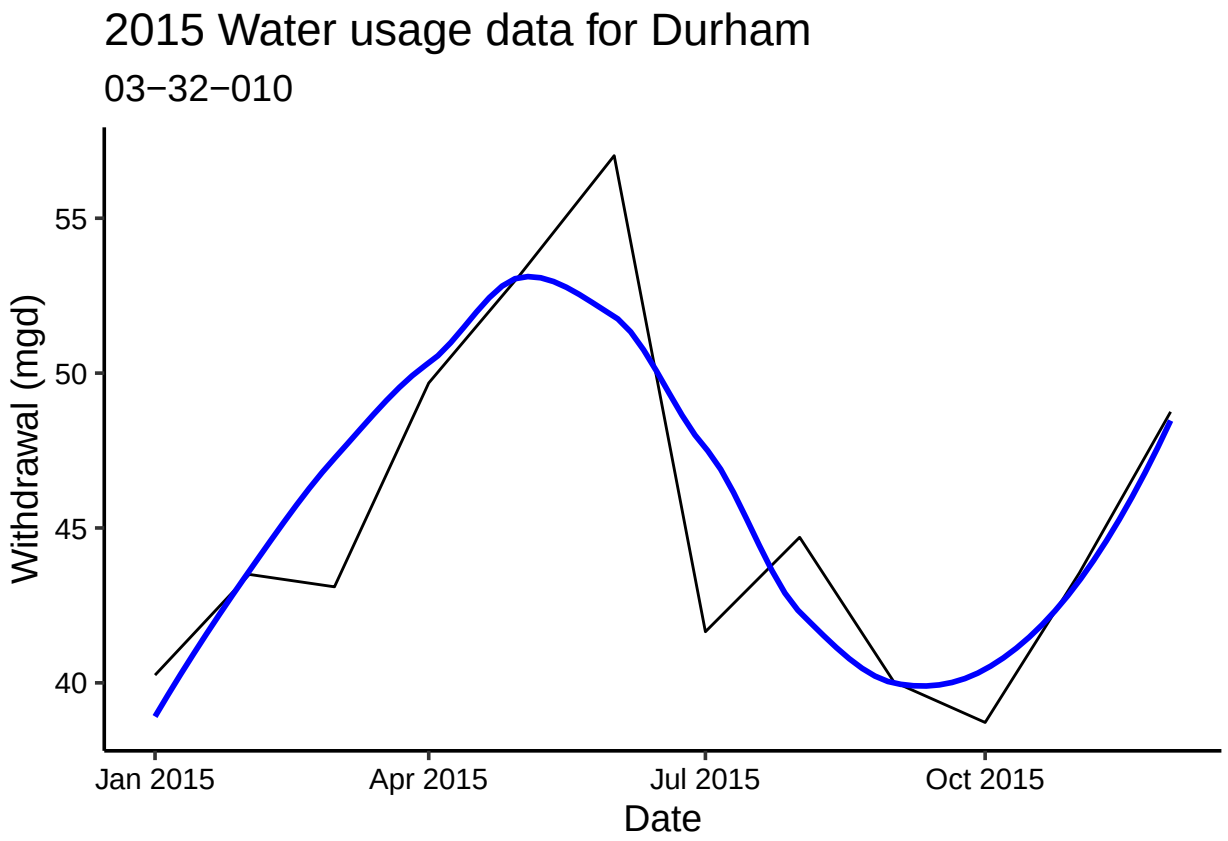
```
#6.
scrape.it <- function(the_pwsid, the_year){
   webpage <- read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP',
                               '/','report.php?pwsid=',the_pwsid, '&year=',the_year))
   water.system.name_tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'
   pwsid_tag <- 'td tr:nth-child(1) td:nth-child(5)'
   ownership_tag <- 'div+ table tr:nth-child(2) td:nth-child(4)'
   max.withdrawals.mgd_tag <- 'th~ td+ td'

   water.system.name <- webpage %>%
     html_nodes(water.system.name_tag)%>%
     html_text()
   pwsid <- webpage %>%
```

```
    html_nodes(pwsid_tag)%>%
    html_text()
  ownership <- webpage %>%
    html_nodes(ownership_tag)%>%
    html_text()
  max.withdrawals.mgd <- webpage %>%
    html_nodes(max.withdrawals.mgd_tag) %>%
    html_text()

  df_withdrawals <- data.frame("Year" = rep(the_year,12),
                               "Month" = c(1,5,9,2,6,10,3,7,11,4,8,12),
                               "max.withdrawals.mgd" = as.numeric(max.withdrawals.mgd)) %>%
    mutate(water.system.name = !!water.system.name,
        pwsid = !!pwsid,
        ownership = !!ownership,
        Date = my(paste(Month,"-",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

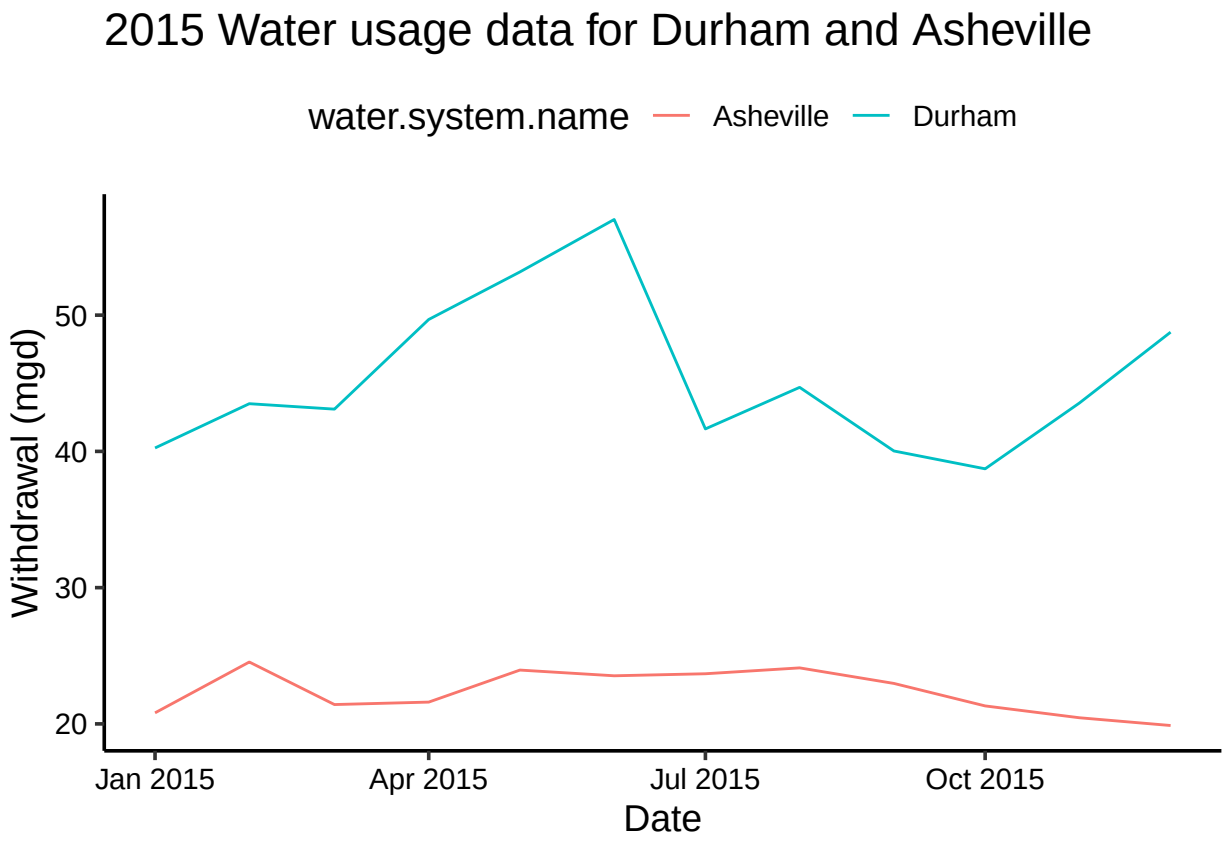
```
#7
Durham.2015_df <- scrape.it('03-32-010',2015)
view(Durham.2015_df)
ggplot(Durham.2015_df,aes(x=Date,y=max.withdrawals.mgd)) +
  geom_line(aes(group=1)) +
  geom_smooth(method="loess",se=FALSE, color="blue") +
  labs(title = paste("2015 Water usage data for",water.system.name),
       subtitle = '03-32-010',
       y="Withdrawal (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 Asheville's to Durham's water withdrawals.

```
#8
Asheville.2015_df <- scrape.it('01-11-010',2015)
view(Asheville.2015_df)
combined_df <- rbind(Durham.2015_df, Asheville.2015_df)
ggplot(combined_df,aes(x=Date,y=max.withdrawals.mgd,color=water.system.name))+
  geom_line() +
  labs(title = paste("2015 Water usage data for Durham and Asheville"),
       y="Withdrawal (mgd)",
       x="Date")
```

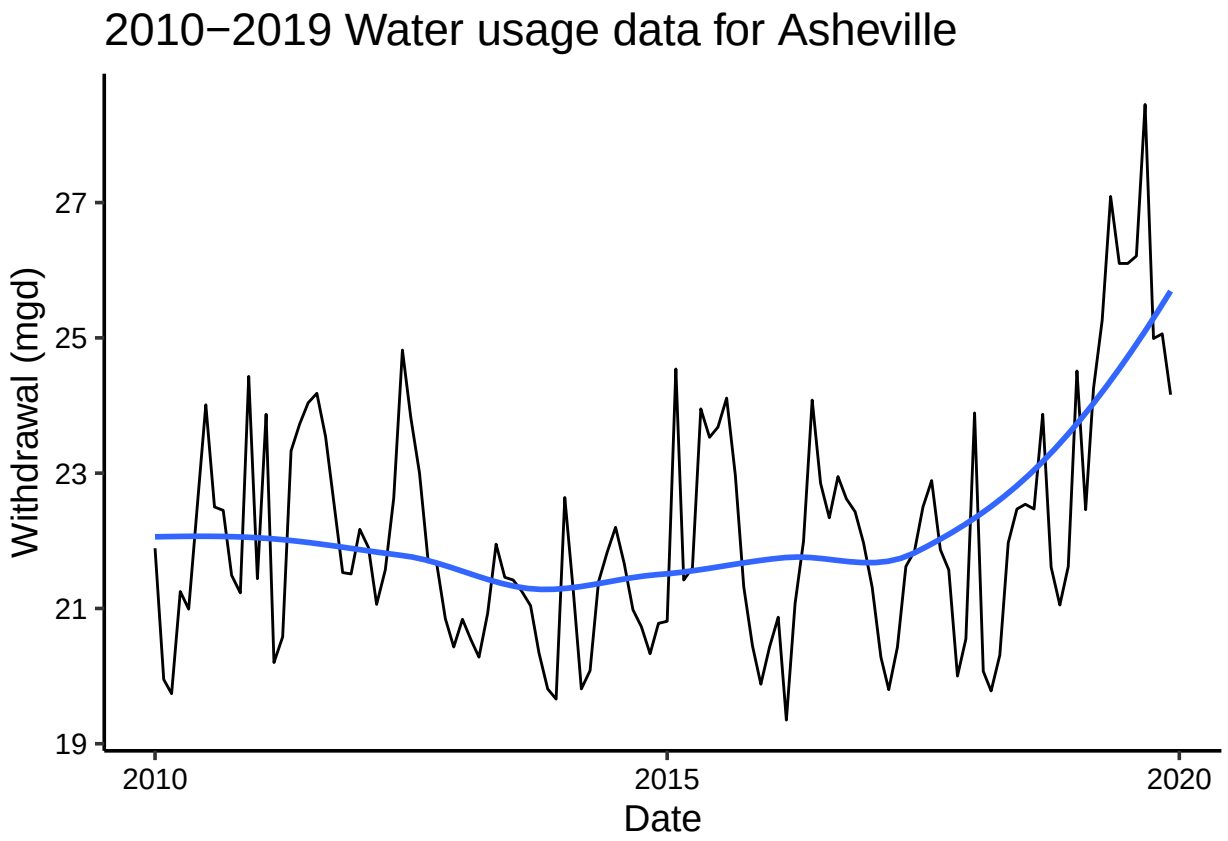


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

   TIP: See Section 3.2 in the "09_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
the_years= c(2010:2019)
the_df <- cross2('01-11-010', the_years) %>%
  map(lift(scrape.it)) %>%
  bind_rows()
ggplot(the_df,aes(x=Date,y=max.withdrawals.mgd)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = "2010-2019 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?

Answer: Yes, generally, Asheville has a upward trend in water usage over time.