Assignment 09: Data Scraping

Yinsu Wang

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/wangyinsu/Desktop/2022 Spring/Env872/Environmental_Data_Analytics_2022"
```

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


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="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.

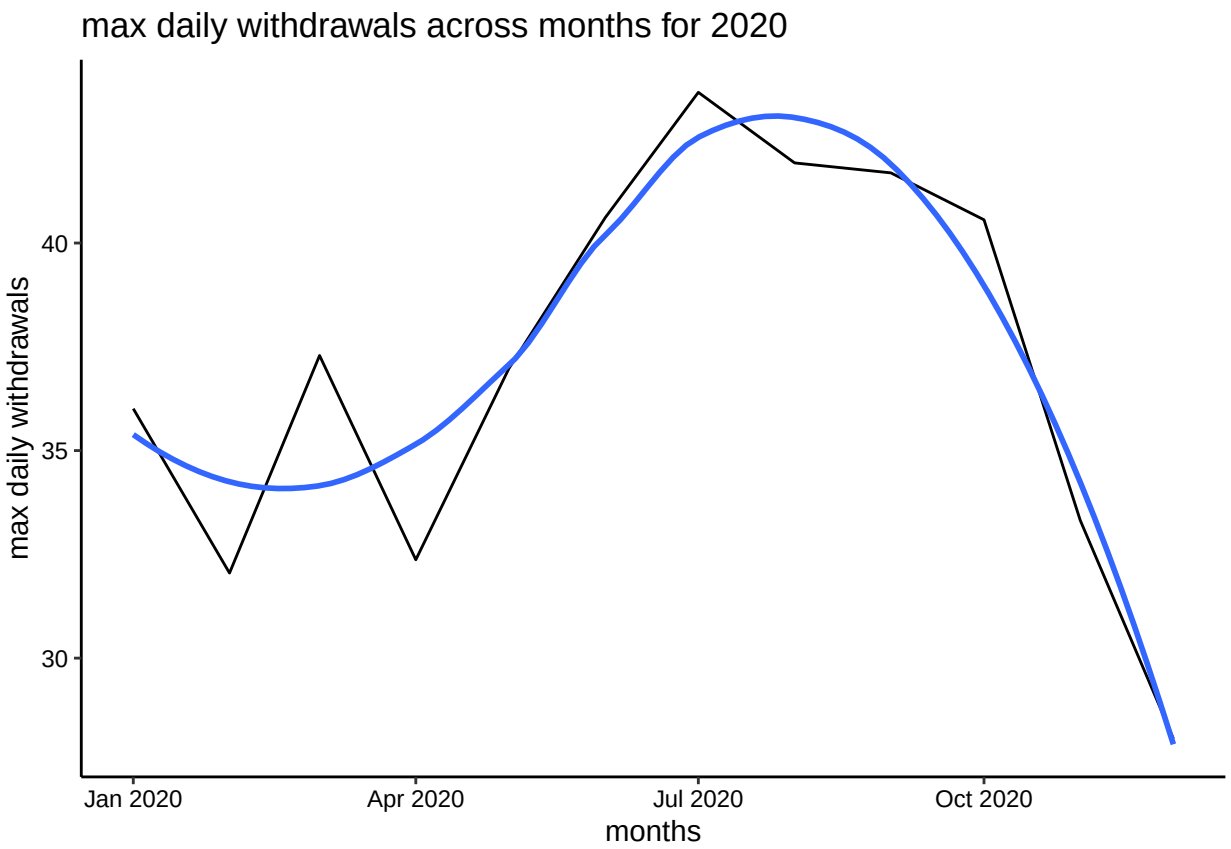
NOTE: It's likely you won't be able to scrape the monthly widthrawal data in order. You can overcome this by creating a month column in the same order the data are scraped: Jan, May, Sept, Feb, etc...

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

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

#5
mgd.plot<-ggplot(data=mgd.scraped,aes(x = Date, y =max.withdrawals.mgd),
                 color = "light blue")+
  geom_line()+
  geom_smooth(method="loess",se=FALSE)+
  labs(y="max daily withdrawals",x="months",
       title = "max daily withdrawals across months for 2020")
mgd.plot
```

```
## `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,Year){
  webpage <- read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',
                              PWSID, '&year=', Year))
 water.system.name <- webpage %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()
pwsid <- 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()
month <- c(1,5,9,2,6,10,3,7,11,4,8,12)
year<-rep(Year,12)
Date<- my(paste(month,"-",year))
df_basic_info<-data.frame(water.system.name=rep(water.system.name),
                          pwsid=rep(PWSID),ownership=rep(ownership),
                          max.withdrawals.mgd=as.numeric(max.withdrawals.mgd),
                          Date)
```

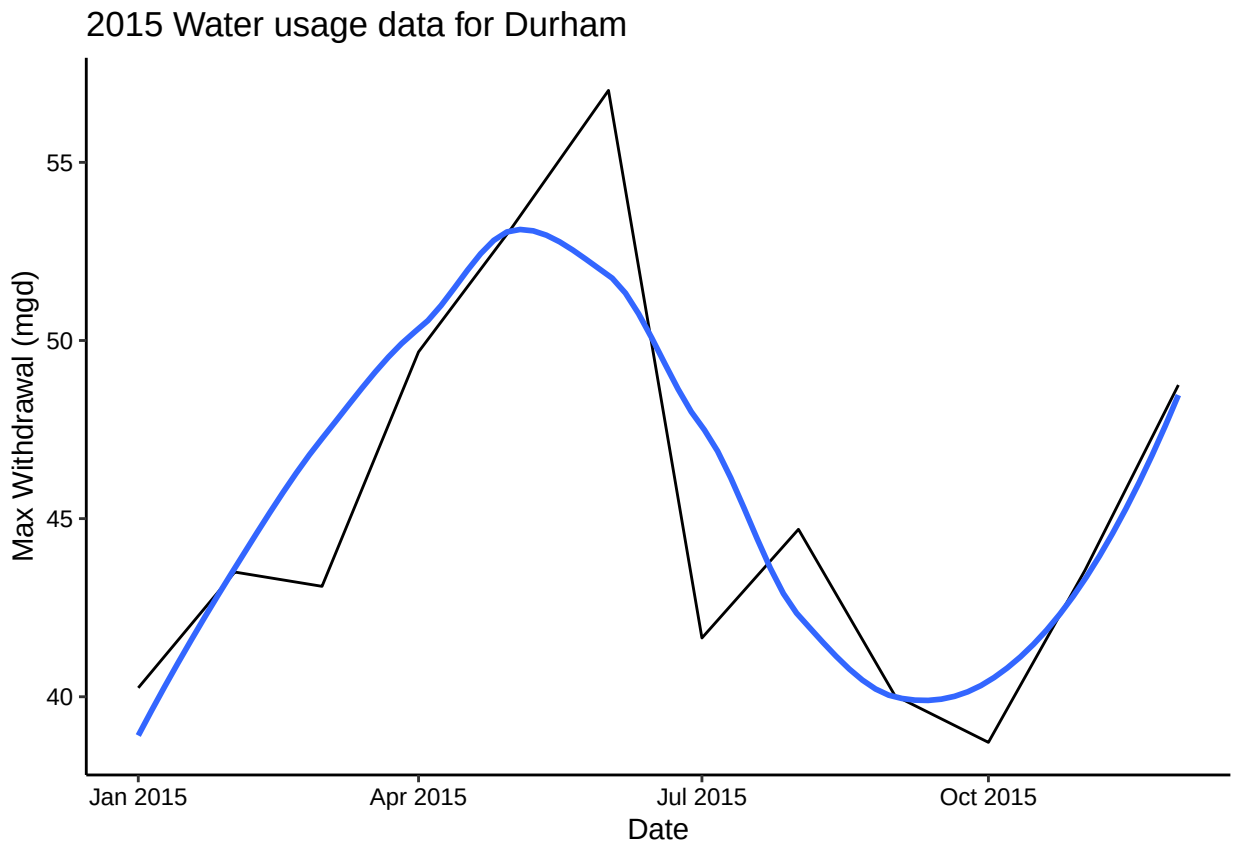
```
  return(df_basic_info)
}
```

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<-scrape.it("03-32-010",2015)

ggplot(Durham.2015,aes(x=Date,y=max.withdrawals.mgd)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2015 Water usage data for",Durham.2015$water.system.name),
       subtitle = Durham.2015$PWSID,
       y="Max Withdrawal (mgd)",
       x="Date")
```

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



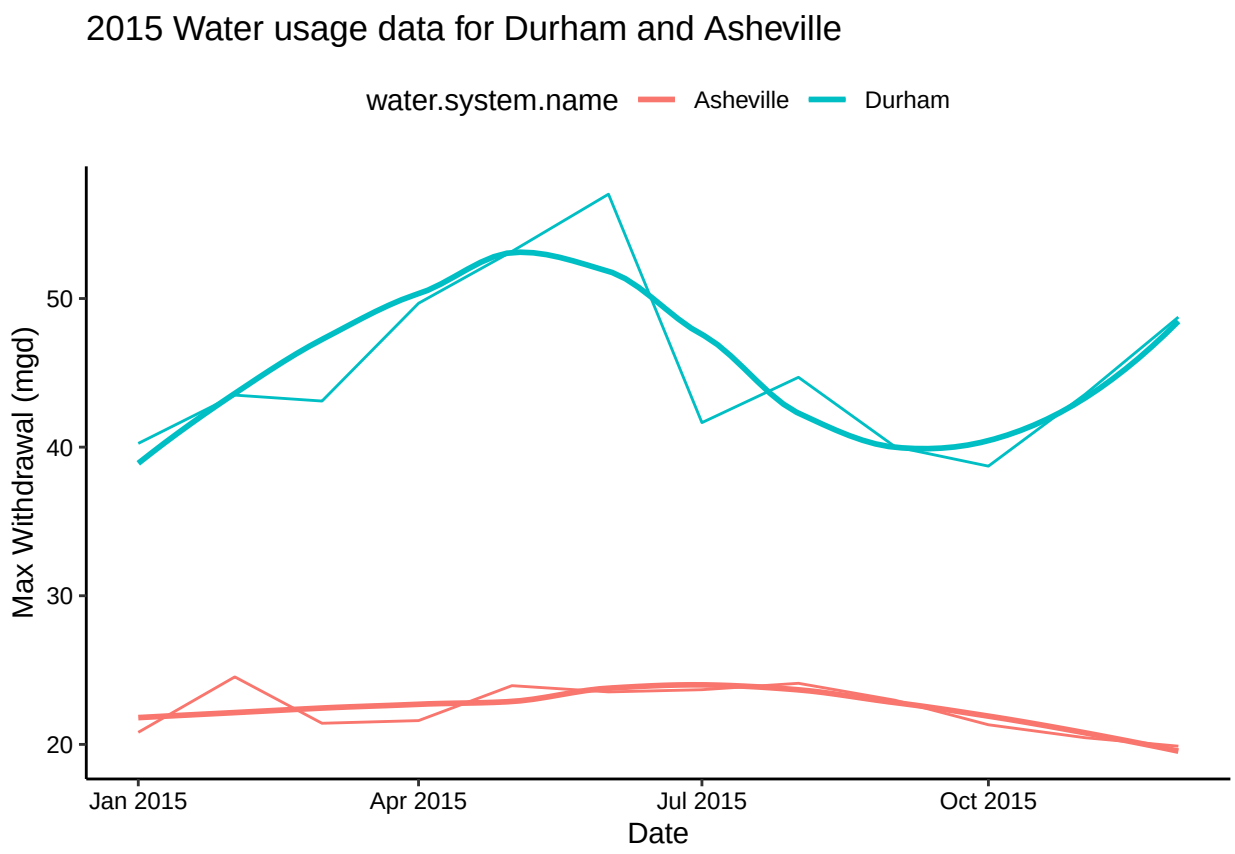
8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares the Asheville to Durham's water withdrawals.

```
#8
Asheville.2015<-scrape.it("01-11-010",2015)
combined.2015<-rbind(Durham.2015,Asheville.2015)

ggplot(combined.2015,aes(x=Date,y=max.withdrawals.mgd,color=water.system.name)) +
```

```
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2015 Water usage data for Durham and Asheville"),
       y="Max Withdrawal (mgd)",
       x="Date")
```

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

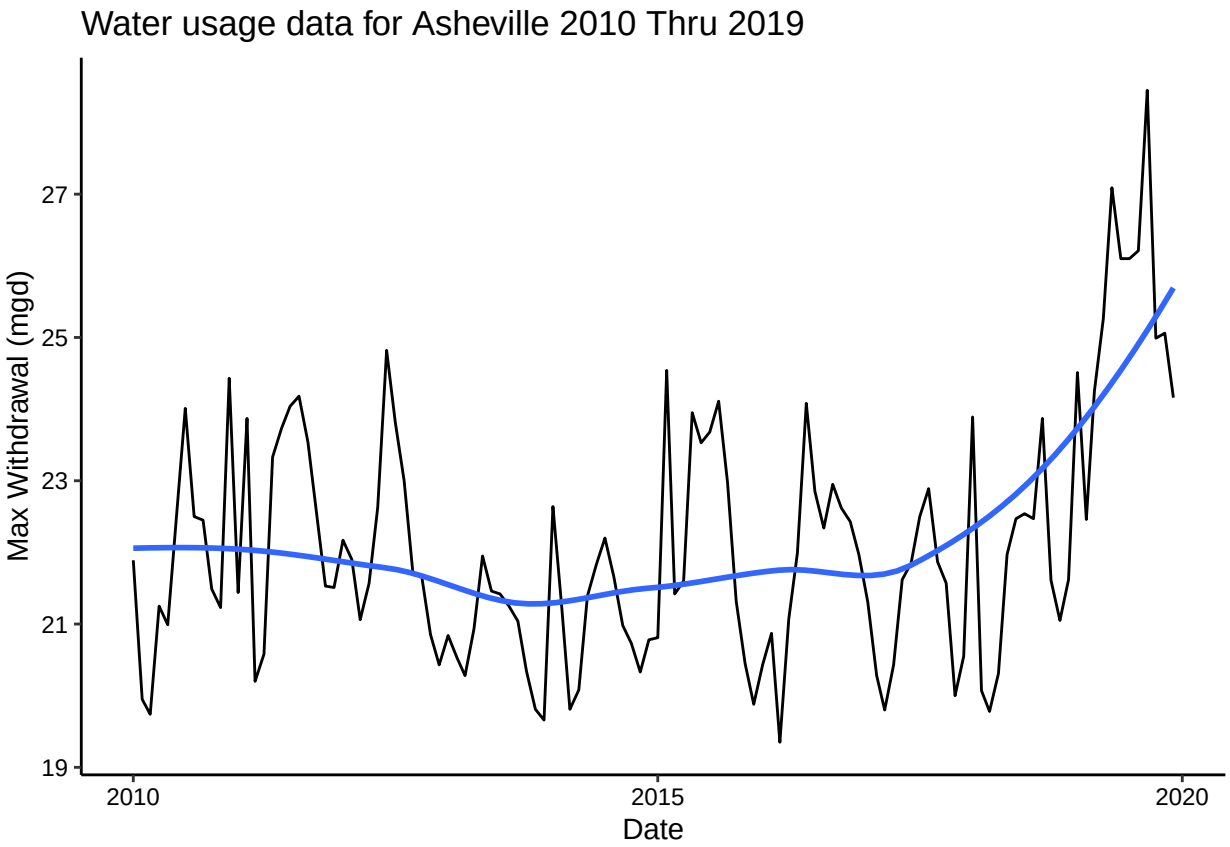


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

```
#9
Asheville.2010<-scrape.it("01-11-010",2010)
Asheville.2011<-scrape.it("01-11-010",2011)
Asheville.2012<-scrape.it("01-11-010",2012)
Asheville.2013<-scrape.it("01-11-010",2013)
Asheville.2014<-scrape.it("01-11-010",2014)
Asheville.2015<-scrape.it("01-11-010",2015)
Asheville.2016<-scrape.it("01-11-010",2016)
Asheville.2017<-scrape.it("01-11-010",2017)
Asheville.2018<-scrape.it("01-11-010",2018)
Asheville.2019<-scrape.it("01-11-010",2019)
Asheville.combined<-rbind(Asheville.2010,Asheville.2011,
                          Asheville.2012,Asheville.2013,Asheville.2014,
                          Asheville.2015,Asheville.2016,Asheville.2017,
                          Asheville.2018,Asheville.2019)
```

```
ggplot(Asheville.combined,aes(x=Date,y=max.withdrawals.mgd)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("Water usage data for Asheville 2010 Thru 2019"),
       y="Max 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? Yes. From the plot we can see a high increase in water usage over time from 2010 thru 2019.