Assignment 09: Data Scraping

Suad Muradov

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/smuradov2021/Desktop/Desktop/MY DOCUMENTS/iMEP/Semester 4/ENV872/Environmental_Data_Analy
```

```
library(tidyverse)
library(lubridate)
library(rvest)
library(ggplot2)

own.theme<-theme_classic(base_size = 18) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "right", legend.title = element_text(size = 14))
theme_set(own.theme)
```

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 fetching the data using URL

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
- PWSID
- Ownership
- From the "3. Water Supply Sources" section:
- MAX 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.2020 <- webpage%>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)")%>%
  html_text()
water.system.name.2020
```

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

```
pwsid.2020 <- webpage%>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)")%>%
  html_text()
pwsid.2020
```

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

```
ownership.2020 <- webpage%>%
  html_nodes('div+ table tr:nth-child(2) td:nth-child(4)')%>%
  html_text()
ownership.2020
```

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

```
max.withdrawals.mgd.2020 <- webpage%>%
  html_nodes("th~ td+ td")%>%
  html_text()
max.withdrawals.mgd.2020
```

```
##  [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...

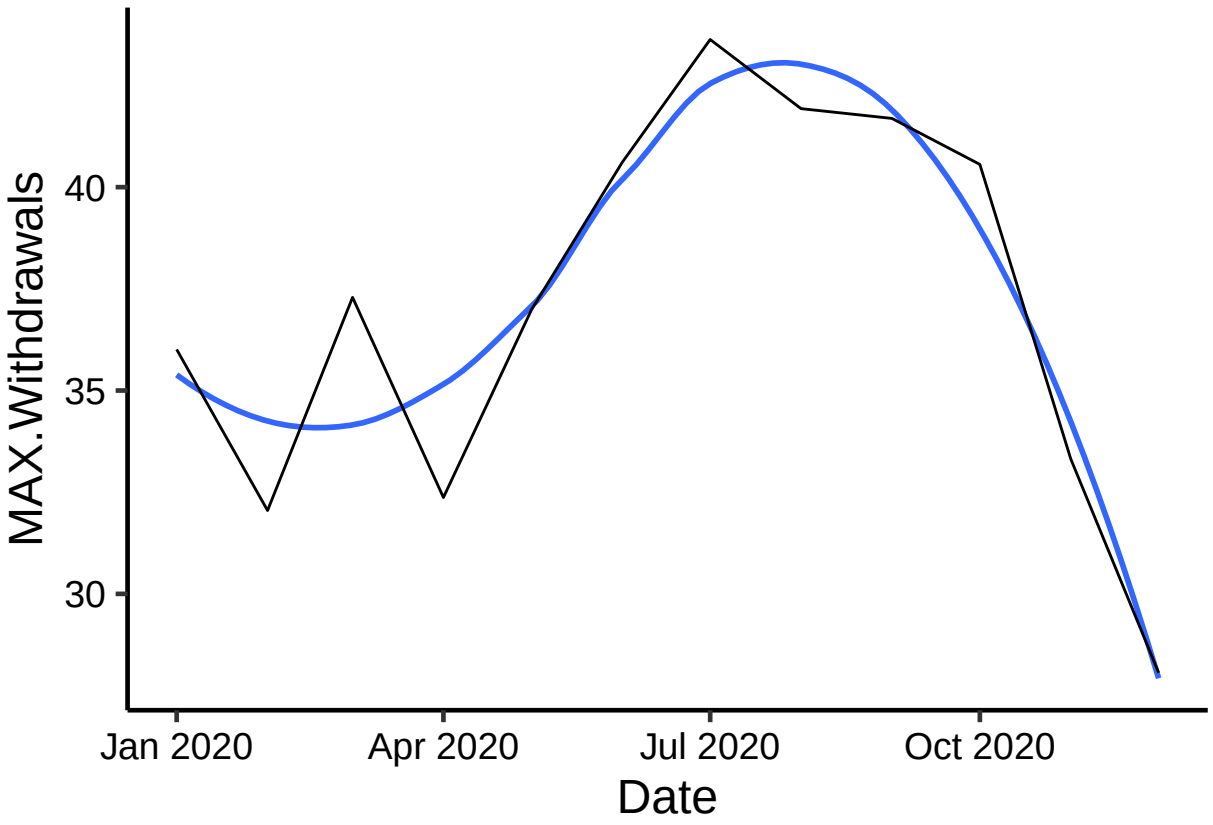
5. Plot the max daily withdrawals across the months for 2020

```
#4
df_withdrawals_2020<-data.frame("Month"=c(1,5,9,2,6,10,3,7,11,4,8,12),
                                "Year"=rep(2020,12),
                                "MAX Withdrawals"=as.numeric(max.withdrawals.mgd.2020))%>%
  mutate(Water_System_Name=!!water.system.name.2020,
         PWSID=!!pwsid.2020,
         Date = my(paste(Month,"-",Year)))%>%
  arrange(Date)
df_withdrawals_2020
```

```
##    Month Year MAX.Withdrawals Water_System_Name     PWSID       Date
## 1      1 2020           36.01            Durham 03-32-010 2020-01-01
## 2      2 2020           32.05            Durham 03-32-010 2020-02-01
## 3      3 2020           37.29            Durham 03-32-010 2020-03-01
## 4      4 2020           32.37            Durham 03-32-010 2020-04-01
## 5      5 2020           36.98            Durham 03-32-010 2020-05-01
## 6      6 2020           40.61            Durham 03-32-010 2020-06-01
## 7      7 2020           43.63            Durham 03-32-010 2020-07-01
## 8      8 2020           41.93            Durham 03-32-010 2020-08-01
## 9      9 2020           41.69            Durham 03-32-010 2020-09-01
## 10    10 2020           40.56            Durham 03-32-010 2020-10-01
## 11    11 2020           33.32            Durham 03-32-010 2020-11-01
## 12    12 2020           28.06            Durham 03-32-010 2020-12-01
```

```
#5
ggplot(df_withdrawals_2020,aes(x=Date,y=MAX.Withdrawals)) +
  geom_smooth(method="loess", se=FALSE)+
  geom_line()
```

```
## `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_year, the_pwsid){

  #Get the proper url
  the_url <- paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',the_pwsid,'&year=',the_ye
print(the_url)

  the_website<-read_html(the_url)
  water.system.name.scraped<- the_website%>%
    html_nodes('div+ table tr:nth-child(1) td:nth-child(2)')%>%
    html_text()
  pwsid.scraped <- the_website%>%
    html_nodes('td tr:nth-child(1) td:nth-child(5)')%>%
    html_text()
  ownership.scraped <- the_website%>%
    html_nodes('div+ table tr:nth-child(2) td:nth-child(4)')%>%
    html_text()
  max.withdrawals.mgd.scraped <- the_website%>%
    html_nodes('th~ td+ td')%>%
    html_text()

  scraped.df<-data.frame(
    "Month" = c(1,5,9,2,6,10,3,7,11,4,8,12),
    "Year" = rep(the_year,12),
```

```
    'Max Daily Withdrawals' = as.numeric(max.withdrawals.mgd.scraped),
    "System.Name" = rep(water.system.name.scraped,12),
    "PWSID" = rep(pwsid.scraped,12),
    "Ownership" = rep(ownership.scraped,12))%>%
    mutate(Date=my(paste(Month,"-",Year )))%>%
    arrange(Date)
  return(scraped.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
withdrawals.2015<-scrape_it(2015, '03-32-010')
```

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

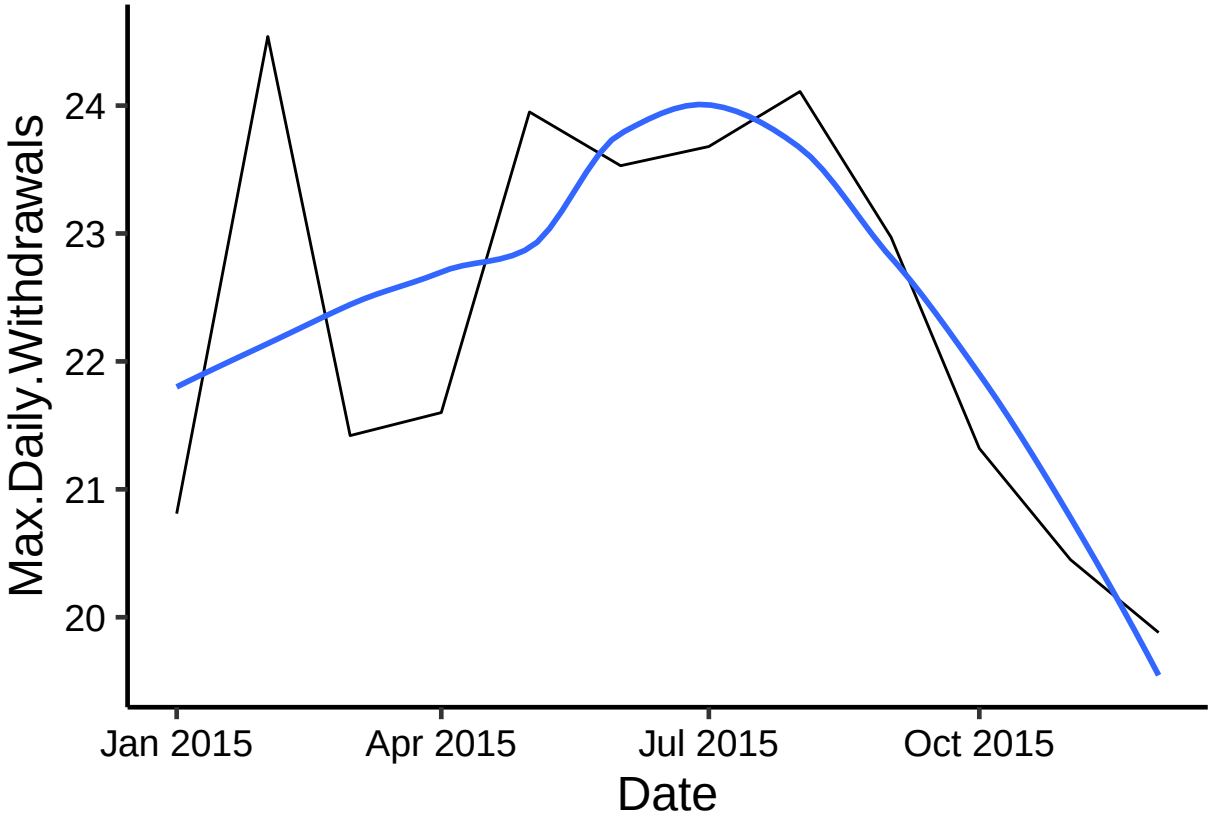
8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares the Asheville to Durham's water withdrawals.

```
#8
withdrawals.ash.2015<-scrape_it(2015, '01-11-010')
```

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

```
ggplot(withdrawals.ash.2015, aes(x=Date, y=Max.Daily.Withdrawals))+
  geom_line()+
  geom_smooth(method="loess", se=FALSE)
```

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



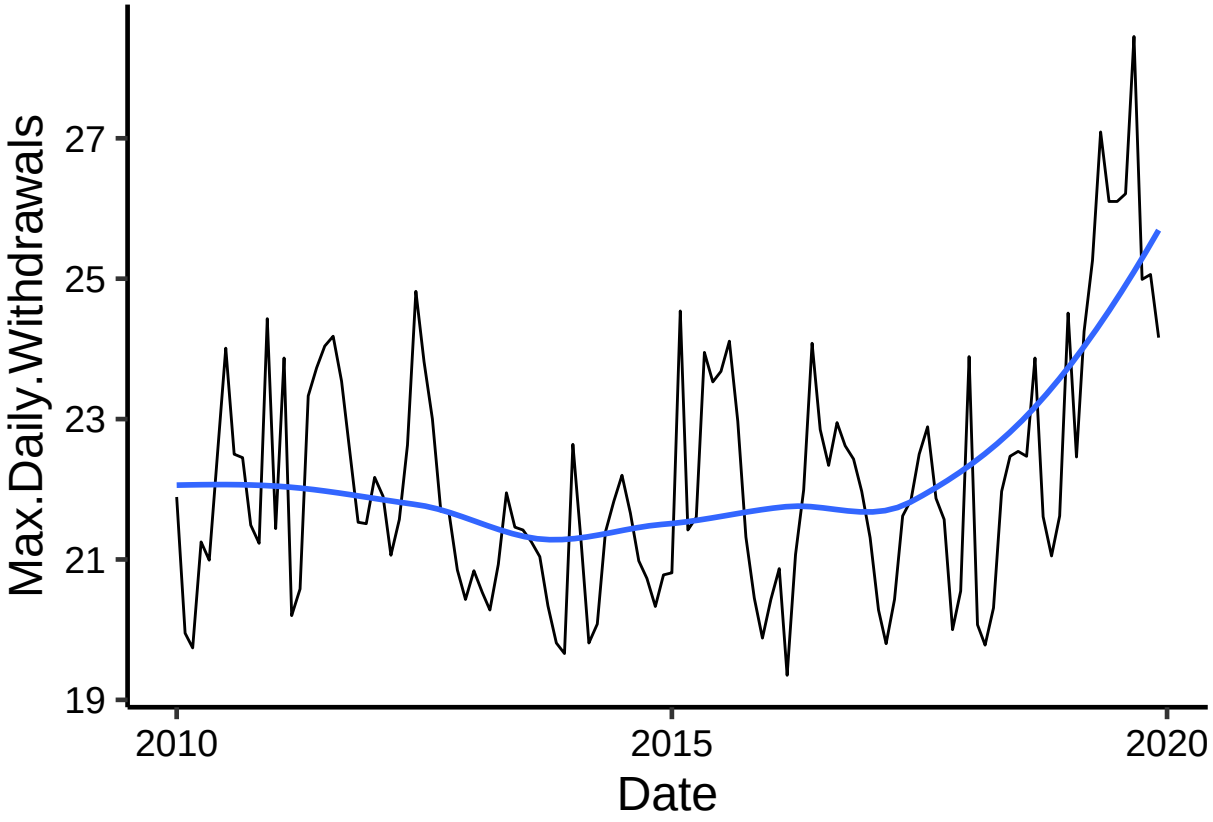
9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

```
#9
withdrawals.ash.2010.19<-map(rep(2010:2019),scrape_it,the_pwsid='01-11-010')%>%
bind_rows()
```

```
## [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"
```

```
ggplot(withdrawals.ash.2010.19, aes(x=Date, y=Max.Daily.Withdrawals))+
  geom_line()+
  geom_smooth(method="loess", se=FALSE)
```

```
## `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? There is an increasing trend in water use