Assignment 10: Data Scraping

Ina Liao

OVERVIEW

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

Directions

1. Rename this file `<FirstLast>_A10_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 your code is tidy; use line breaks to ensure your code fits in the knitted output.
5. Be sure to **answer the questions** in this assignment document.
6. When you have completed the assignment, **Knit** the text and code into a single PDF file.

Set up

1. Set up your session:

- Load the packages `tidyverse`, `rvest`, and any others you end up using.
- Check your working directory

```
#1
library(tidyverse)
library(lubridate)
library(viridis)
library(here)
#install.packages("rvest")
library(rvest)
#install.packages("dataRetrieval")
library(dataRetrieval)
#install.packages("tidycensus")
library(tidycensus)

here()
```

```
## [1] "/Users/inaliao/Desktop/EDE_Fall2023_2"
```

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham's 2022 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=2022

Indicate this website as the as the URL to be scraped. (In other words, read the contents into an `rvest` webpage object.)

```
#2
theURL <-'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2022'
website<-read_html(theURL)
```

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:
- Maximum Day 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)".

```
#3
water_system_name<-website %>%
  html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>%
  html_text()

PWSID<-website %>%
  html_nodes('td tr:nth-child(1) td:nth-child(5)') %>%
  html_text()

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

max_day_use_1234<-website %>%
  html_nodes('th~ td:nth-child(3)') %>%
  html_text()
max_day_use_5678<-website %>%
  html_nodes('th~ td+ td:nth-child(6)')%>%
  html_text()
max_day_use_9101112<-website %>%
  html_nodes('th~ td+ td:nth-child(9)')%>%
  html_text()

#change data type
max_day_use<-cbind(max_day_use_1234,max_day_use_5678,max_day_use_9101112)
max_day_use<-as.numeric(max_day_use)
```

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... Or, you could scrape month values from the web page...

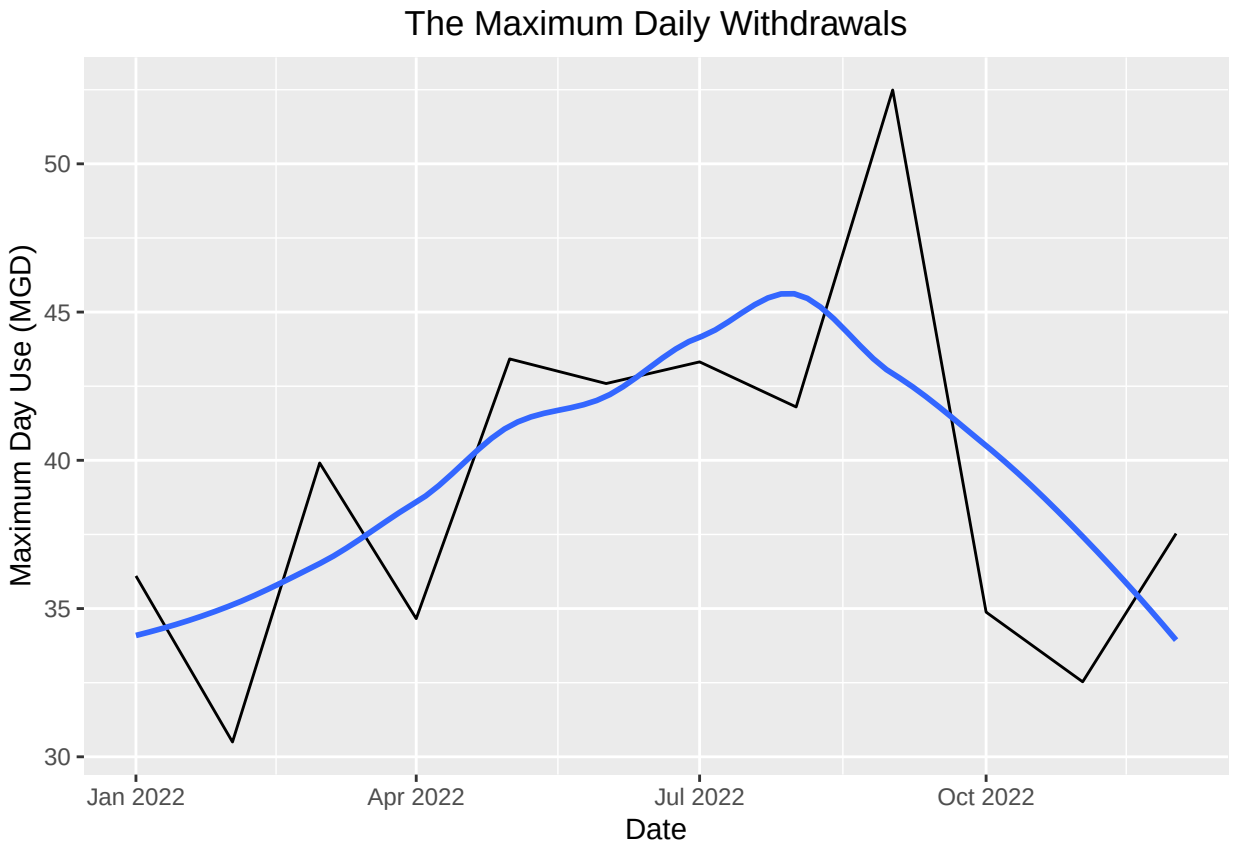
5. Create a line plot of the maximum daily withdrawals across the months for 2022

```
#4 convert scraped data into a dataframe
df_water<-data.frame('Month'=rep(1:12),
                     'Year'=2022,
                     'Water_System_Name'=water_system_name,
                     'PWSID'=PWSID,
                     'Ownership'=ownership,
                     'Maximum_Day_Use_MGD'=max_day_use)%>%
  mutate(Date=my(paste(Month,"-",Year))) #add another Date column that includes month and year
head(df_water)
```

```
##   Month Year Water_System_Name     PWSID    Ownership Maximum_Day_Use_MGD
## 1     1 2022            Durham 03-32-010 Municipality               36.10
## 2     2 2022            Durham 03-32-010 Municipality               30.50
## 3     3 2022            Durham 03-32-010 Municipality               39.91
## 4     4 2022            Durham 03-32-010 Municipality               34.66
## 5     5 2022            Durham 03-32-010 Municipality               43.42
## 6     6 2022            Durham 03-32-010 Municipality               42.59
##         Date
## 1 2022-01-01
## 2 2022-02-01
## 3 2022-03-01
## 4 2022-04-01
## 5 2022-05-01
## 6 2022-06-01
```

```
#5 create a line plot
ggplot(df_water,aes(x=Date,y=Maximum_Day_Use_MGD))+
  geom_line()+
  geom_smooth(method="loess",se=FALSE) +
  labs(title='The Maximum Daily Withdrawals',
       y='Maximum Day Use (MGD)',
       x='Date')+
  theme(plot.title=element_text(hjust=0.5))
```

```
## `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. construct a function to scrape data for any PWSID and year (inputs: PWSID and year)

scrape1<-function(the_PWSID,the_Year){
  # retrieve the website contents
  the_website<-paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',the_PWSID,'&year=',the_
  THE_HTML<-read_html(the_website)

  # scrape the data items (refer to "the_website")
  the_water_system_name<-THE_HTML %>%
    html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>%
    html_text()

  the_PWSID<-THE_HTML %>%
    html_nodes('td tr:nth-child(1) td:nth-child(5)') %>%
    html_text()

  the_ownership<-THE_HTML %>%
    html_nodes('div+ table tr:nth-child(2) td:nth-child(4)') %>%
    html_text()

  the_max_day_use_1234<-THE_HTML %>%
    html_nodes('th~ td:nth-child(3)') %>%
    html_text()
  the_max_day_use_5678<-THE_HTML%>%
```

```
    html_nodes('th~ td+ td:nth-child(6)')%>%
    html_text()
  the_max_day_use_9101112<-THE_HTML%>%
    html_nodes('th~ td+ td:nth-child(9)')%>%
    html_text()
  the_max_day_use<-cbind(the_max_day_use_1234,the_max_day_use_5678,the_max_day_use_9101112)

  # covert to a dataframe
  df_water_function<-data.frame('Month'=rep(1:12),
                                'Year'=rep(the_Year,12),
                                'Maximum_Day_Use_MGD'=as.numeric(the_max_day_use),
                                'Water_System_Name'=the_water_system_name,
                                'PWSID'=the_PWSID,
                                'Ownership'=the_ownership)%>%
    mutate(Date=my(paste(Month,"-",Year)))

  # return the dataframe
  return(df_water_function)
}
```

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

```
#7
df_Durham<-scrape1('03-32-010',2015)
df_Durham
```

```
##    Month Year Maximum_Day_Use_MGD Water_System_Name     PWSID    Ownership
## 1      1 2015               40.25            Durham 03-32-010 Municipality
## 2      2 2015               43.50            Durham 03-32-010 Municipality
## 3      3 2015               43.10            Durham 03-32-010 Municipality
## 4      4 2015               49.68            Durham 03-32-010 Municipality
## 5      5 2015               53.17            Durham 03-32-010 Municipality
## 6      6 2015               57.02            Durham 03-32-010 Municipality
## 7      7 2015               41.65            Durham 03-32-010 Municipality
## 8      8 2015               44.70            Durham 03-32-010 Municipality
## 9      9 2015               40.03            Durham 03-32-010 Municipality
## 10    10 2015               38.72            Durham 03-32-010 Municipality
## 11    11 2015               43.55            Durham 03-32-010 Municipality
## 12    12 2015               48.75            Durham 03-32-010 Municipality
##          Date
## 1  2015-01-01
## 2  2015-02-01
## 3  2015-03-01
## 4  2015-04-01
## 5  2015-05-01
## 6  2015-06-01
## 7  2015-07-01
## 8  2015-08-01
## 9  2015-09-01
## 10 2015-10-01
## 11 2015-11-01
## 12 2015-12-01
```

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.

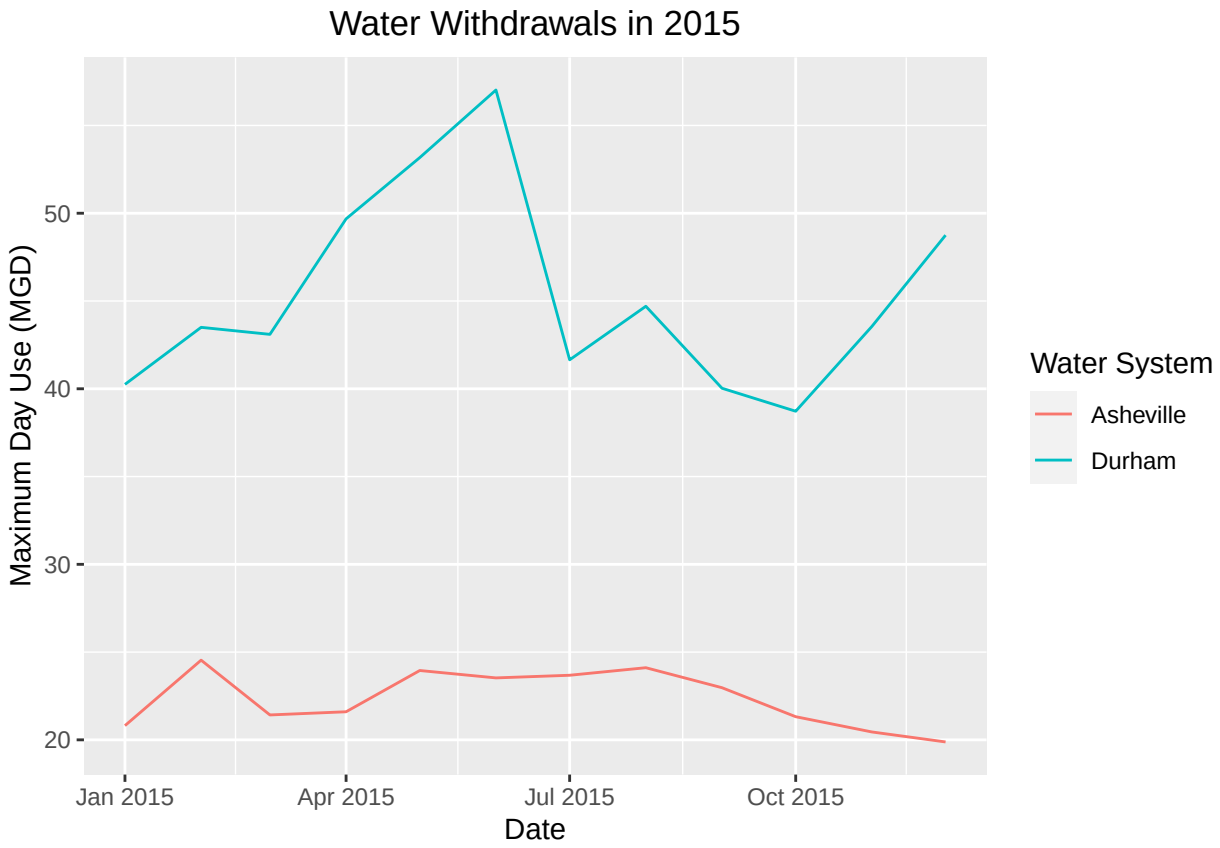
```
#8
df_Asheville<-scrape1('01-11-010',2015)
df_Asheville
```

```
##    Month Year Maximum_Day_Use_MGD Water_System_Name     PWSID    Ownership
## 1      1 2015               20.81         Asheville 01-11-010 Municipality
## 2      2 2015               24.54         Asheville 01-11-010 Municipality
## 3      3 2015               21.42         Asheville 01-11-010 Municipality
## 4      4 2015               21.60         Asheville 01-11-010 Municipality
## 5      5 2015               23.95         Asheville 01-11-010 Municipality
## 6      6 2015               23.53         Asheville 01-11-010 Municipality
## 7      7 2015               23.68         Asheville 01-11-010 Municipality
## 8      8 2015               24.11         Asheville 01-11-010 Municipality
## 9      9 2015               22.97         Asheville 01-11-010 Municipality
## 10    10 2015               21.32         Asheville 01-11-010 Municipality
## 11    11 2015               20.45         Asheville 01-11-010 Municipality
## 12    12 2015               19.88         Asheville 01-11-010 Municipality
##          Date
## 1  2015-01-01
## 2  2015-02-01
## 3  2015-03-01
## 4  2015-04-01
## 5  2015-05-01
## 6  2015-06-01
## 7  2015-07-01
## 8  2015-08-01
## 9  2015-09-01
## 10 2015-10-01
## 11 2015-11-01
## 12 2015-12-01
```

```
# combine two dataframes
df_Dur_Ash<-rbind(df_Durham, df_Asheville)
colnames(df_Dur_Ash)
```

```
## [1] "Month"               "Year"                "Maximum_Day_Use_MGD"
## [4] "Water_System_Name"   "PWSID"               "Ownership"
## [7] "Date"
```

```
# create a plot
ggplot(df_Dur_Ash,aes(x=Date,y=Maximum_Day_Use_MGD,color=Water_System_Name))+
  geom_line()+
  labs(title='Water Withdrawals in 2015',
       x='Date',
       y='Maximum Day Use (MGD)',
       color='Water System')+
  theme(plot.title = element_text(hjust = 0.5))
```



9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2021.Add a smoothed line to the plot (method = 'loess').

   TIP: See Section 3.2 in the "10_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

# Asheville water system
the_PWSID<-'01-11-010'

# create a sequence from 2010 to 2021
the_Year_10_21<-rep(2010:2021)

# use the function to grab values into a list of dataframe
df_Ash_2010_2021<- map2(the_PWSID,the_Year_10_21,scrape1)  %>%
  bind_rows()
head(df_Ash_2010_2021)
```

```
##   Month Year Maximum_Day_Use_MGD Water_System_Name     PWSID    Ownership
## 1     1 2010               21.89         Asheville 01-11-010 Municipality
## 2     2 2010               19.95         Asheville 01-11-010 Municipality
## 3     3 2010               19.74         Asheville 01-11-010 Municipality
## 4     4 2010               21.25         Asheville 01-11-010 Municipality
## 5     5 2010               20.99         Asheville 01-11-010 Municipality
## 6     6 2010               22.53         Asheville 01-11-010 Municipality
##         Date
```

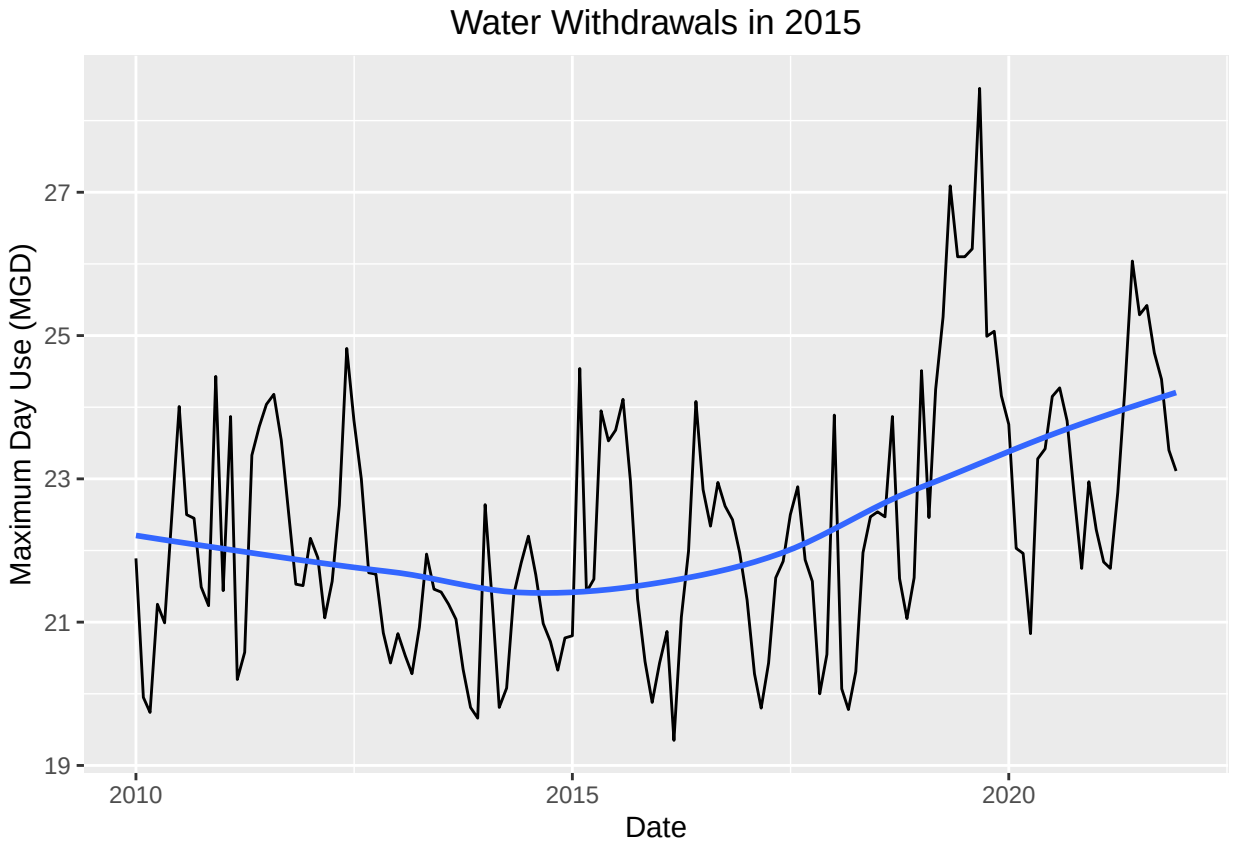
```
## 1 2010-01-01
## 2 2010-02-01
## 3 2010-03-01
## 4 2010-04-01
## 5 2010-05-01
## 6 2010-06-01
```

```
colnames(df_Ash_2010_2021)
```

```
## [1] "Month"                "Year"                 "Maximum_Day_Use_MGD"
## [4] "Water_System_Name"    "PWSID"                "Ownership"
## [7] "Date"
```

```
# plot
ggplot(df_Ash_2010_2021,aes(x=Date,y=Maximum_Day_Use_MGD))+
  geom_line()+
  geom_smooth(method="loess",se=FALSE) +
  labs(title='Water Withdrawals in 2015',
       x='Date',
       y='Maximum Day Use (MGD)',
       color='Water System')+
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `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: The maximum water usage exhibits seasonality, and has demonstrated an upward trend from 2010 to 2021 from the graph. However, the plot can not tell us if total water usage has increased since 2021.