Assignment 10: Data Scraping

Kendra Sultzer

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_10_Data_Scraping.Rmd") prior to submission.

The completed exercise is due on Tuesday, April 6 at 11:59 pm.

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() #looks good
```

```
## [1] "C:/Users/kendr/Documents/Spring2021Classes/ENV872/Environmental_Data_Analytics_2021/Assignments
```

```
library(tidyverse)
```

```
## Warning: package 'tibble' was built under R version 4.0.4
```

```
library(rvest)
library(lubridate)
library(zoo)

mytheme <- theme_classic(base_size = 12) +
```

```
  theme(axis.text = element_text(color = "black"),
        legend.position = "right",
        legend.box.background = element_rect(color="black"), legend.background = element_blank(),
        plot.title=element_text(hjust = 0.5))
#makes font size bigger, puts legend on right, puts box around legend, centers title in middle
theme_set(mytheme) #setting for all subsequent plots
```

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=2019

Indicate this website as the as the URL to be scraped.

```
#2
my_website <- read_html("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2019")
```

3. The data we want to collect are listed below:

- From the "System Information" section:
- Water system name
- PSWID
- Ownership
- From the "Water Supply Sources" section:
- Maximum monthly withdrawals (MGD)

In the code chunk below scrape these values into the supplied variable names.

```
#3
Water_system_name <- my_website %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text
PSWID <- my_website %>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
  html_text
Ownership <- my_website %>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%
  html_text
Max_month_WD <- my_website %>%
  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.)

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...

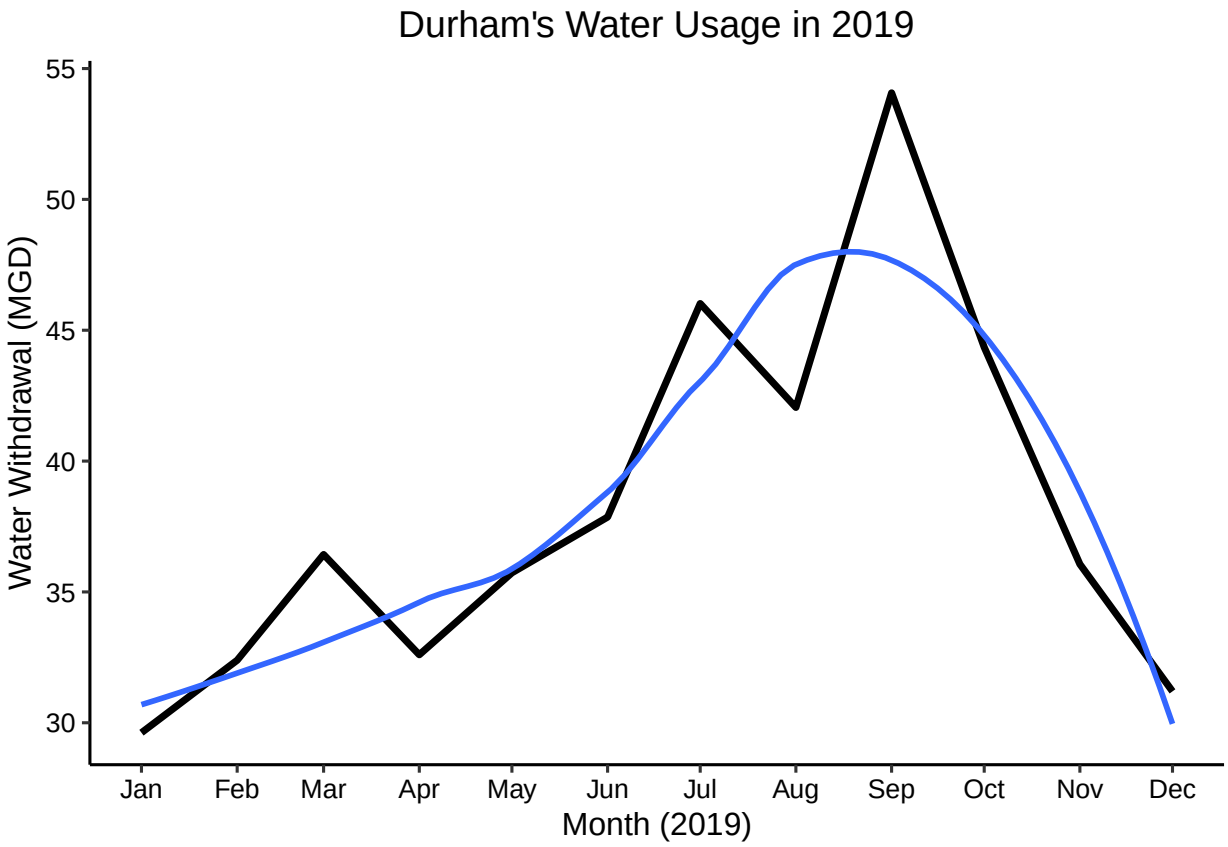
5. Plot the max daily withdrawals across the months for 2019.

```
#4
dataframe <- data.frame(
  Water.System.Name=Water_system_name,
  ID=PSWID,
  Ownership=Ownership,
  Max.monthly.WD=as.numeric(Max_month_WD),
  Month=c("Jan","May", "Sep", "Feb", "Jun", "Oct", "Mar", "Jul", "Nov", "Apr", "Aug", "Dec"),
  Year=("2019"))

dataframe$Date <- as.yearmon(paste(dataframe$Year, dataframe$Month), "%Y %b") #creating date column
dataframe$Date <- as.Date(dataframe$Date, format="%b %Y")#making it manipulative for graphing


#5
ggplot(dataframe, aes(x=Date, y=Max.monthly.WD))+
  geom_line(size=1.25)+
  geom_smooth(method="loess",se=FALSE)+
  labs(title="Durham's Water Usage in 2019",
       y="Water Withdrawal (MGD)",
       x="Month (2019)")+
  scale_x_date(date_labels = "%b", date_breaks = "1 month")
```

```
## `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 data scraped.

```
#6.

scrape.it <- function(the_year, the_pwsid){

  #make the base url
the_base_url <- "https://www.ncwater.org/WUDC/app/LWSP/report.php"
the_scrape_url <- paste0(the_base_url, "?pwsid=", the_pwsid, "&year=", the_year)
the_scrape_url

  #Retrieve the website contents
  the_website <- read_html(the_scrape_url)

  #Scrape the data
Water_system_name <- the_website %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text
PSWID <- the_website %>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
  html_text
Ownership <- the_website %>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%
```

```
  html_text
Max_month_WD <- the_website %>%
  html_nodes("th~ td+ td") %>%
  html_text

  #Convert to dataframe
df_withdrawals <- data.frame(
  Water.System.Name=Water_system_name,
  ID=PSWID,
  Ownership=Ownership,
  Max.monthly.WD=as.numeric(Max_month_WD),
  Month=c("Jan","May", "Sep", "Feb", "Jun", "Oct", "Mar", "Jul", "Nov", "Apr", "Aug", "Dec"),
  Year=(the_year)
)

#creating date column
df_withdrawals$Date <- as.yearmon(paste(df_withdrawals$Year, df_withdrawals$Month), "%Y %b")
#making it manipulative for graphing
df_withdrawals$Date <- as.Date(df_withdrawals$Date, format="%b %Y")

  #Return the dataframe
  return(df_withdrawals)
}
```

7. Use the function above to extract and plot max daily withdrawals for Durham for each month in 2015

```
#7
scrape.it(2015,"03-32-010")
```

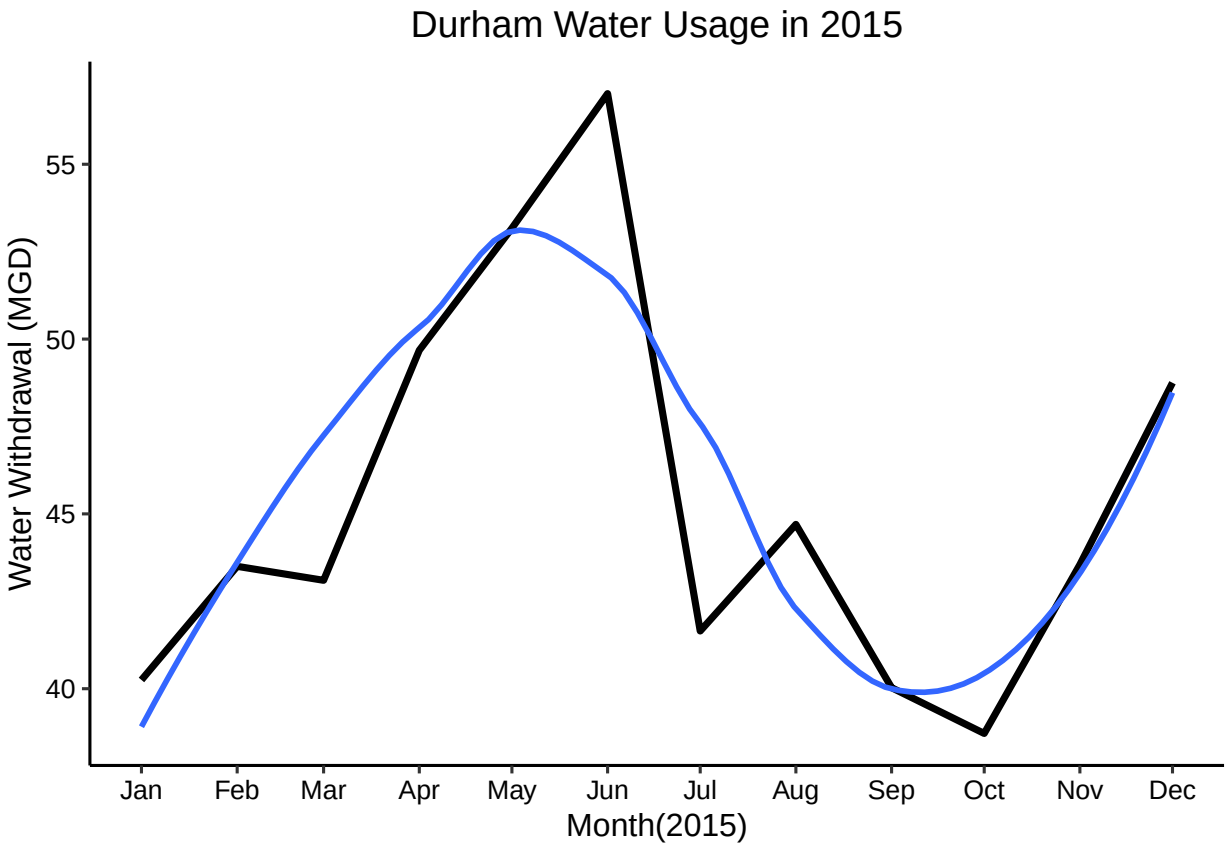
```
##    Water.System.Name        ID    Ownership Max.monthly.WD Month Year
## 1             Durham 03-32-010 Municipality          40.25   Jan 2015
## 2             Durham 03-32-010 Municipality          53.17   May 2015
## 3             Durham 03-32-010 Municipality          40.03   Sep 2015
## 4             Durham 03-32-010 Municipality          43.50   Feb 2015
## 5             Durham 03-32-010 Municipality          57.02   Jun 2015
## 6             Durham 03-32-010 Municipality          38.72   Oct 2015
## 7             Durham 03-32-010 Municipality          43.10   Mar 2015
## 8             Durham 03-32-010 Municipality          41.65   Jul 2015
## 9             Durham 03-32-010 Municipality          43.55   Nov 2015
## 10            Durham 03-32-010 Municipality          49.68   Apr 2015
## 11            Durham 03-32-010 Municipality          44.70   Aug 2015
## 12            Durham 03-32-010 Municipality          48.75   Dec 2015
##          Date
## 1  2015-01-01
## 2  2015-05-01
## 3  2015-09-01
## 4  2015-02-01
## 5  2015-06-01
## 6  2015-10-01
## 7  2015-03-01
## 8  2015-07-01
## 9  2015-11-01
## 10 2015-04-01
```

```
## 11 2015-08-01
## 12 2015-12-01
```

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

#plot
ggplot(Durham.WD, aes(x=Date, y=Max.monthly.WD))+
  geom_line(size=1.25)+
  geom_smooth(method="loess",se=FALSE)+
  labs(title="Durham Water Usage in 2015",
       y="Water Withdrawal (MGD)",
       x="Month(2015)")+
  scale_x_date(date_labels = "%b", date_breaks = "1 month")
```

```
## `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
scrape.it(2015, "01-11-010") #data from Asheville
```

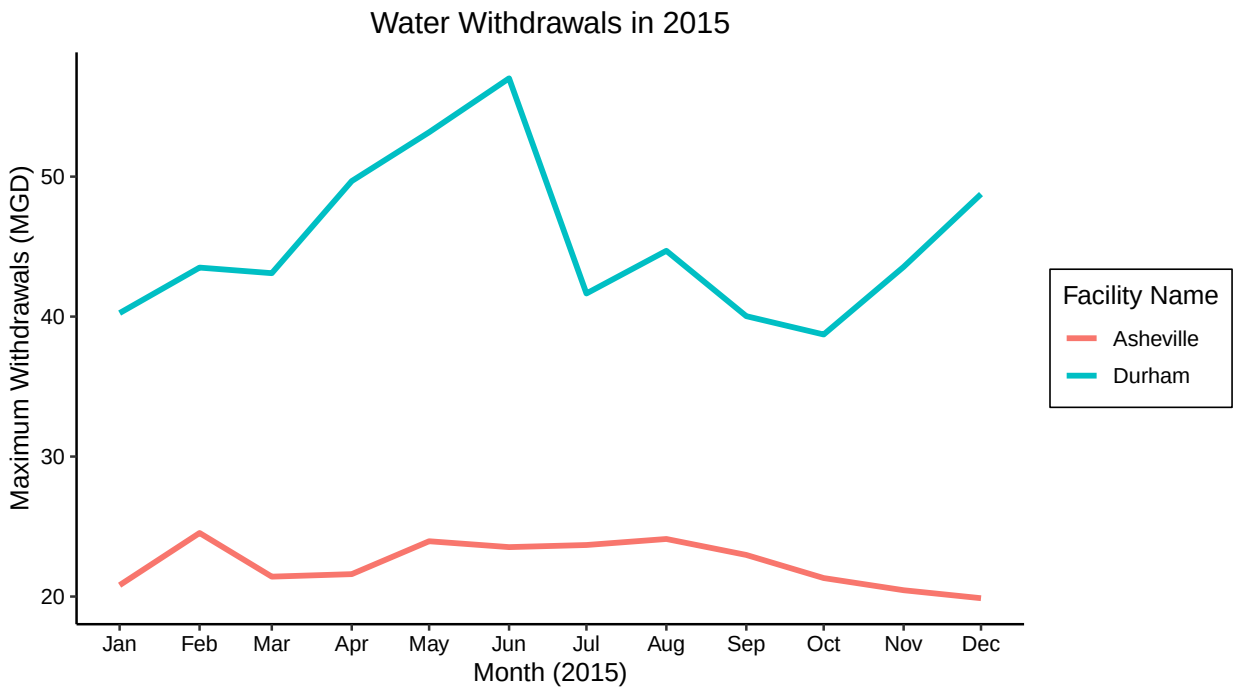
```
##    Water.System.Name        ID    Ownership Max.monthly.WD Month Year
```

```
## 1          Asheville 01-11-010 Municipality          20.81   Jan 2015
## 2          Asheville 01-11-010 Municipality          23.95   May 2015
## 3          Asheville 01-11-010 Municipality          22.97   Sep 2015
## 4          Asheville 01-11-010 Municipality          24.54   Feb 2015
## 5          Asheville 01-11-010 Municipality          23.53   Jun 2015
## 6          Asheville 01-11-010 Municipality          21.32   Oct 2015
## 7          Asheville 01-11-010 Municipality          21.42   Mar 2015
## 8          Asheville 01-11-010 Municipality          23.68   Jul 2015
## 9          Asheville 01-11-010 Municipality          20.45   Nov 2015
## 10         Asheville 01-11-010 Municipality          21.60   Apr 2015
## 11         Asheville 01-11-010 Municipality          24.11   Aug 2015
## 12         Asheville 01-11-010 Municipality          19.88   Dec 2015
##          Date
## 1  2015-01-01
## 2  2015-05-01
## 3  2015-09-01
## 4  2015-02-01
## 5  2015-06-01
## 6  2015-10-01
## 7  2015-03-01
## 8  2015-07-01
## 9  2015-11-01
## 10 2015-04-01
## 11 2015-08-01
## 12 2015-12-01
```

```
Asheville.WD <- scrape.it(2015, "01-11-010") #saving as dataframes

WD_join <- rbind(Durham.WD,Asheville.WD)

#plot comparisons
ggplot(WD_join, aes(x=Date, y=Max.monthly.WD))+
  geom_line(aes(color=Water.System.Name), size=1.25)+
  labs(title="Water Withdrawals in 2015",
       y="Maximum Withdrawals (MGD)",
       x="Month (2015)")+
  scale_color_discrete(name="Facility Name")+
  scale_x_date(date_labels = "%b", date_breaks = "1 month")
```



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.

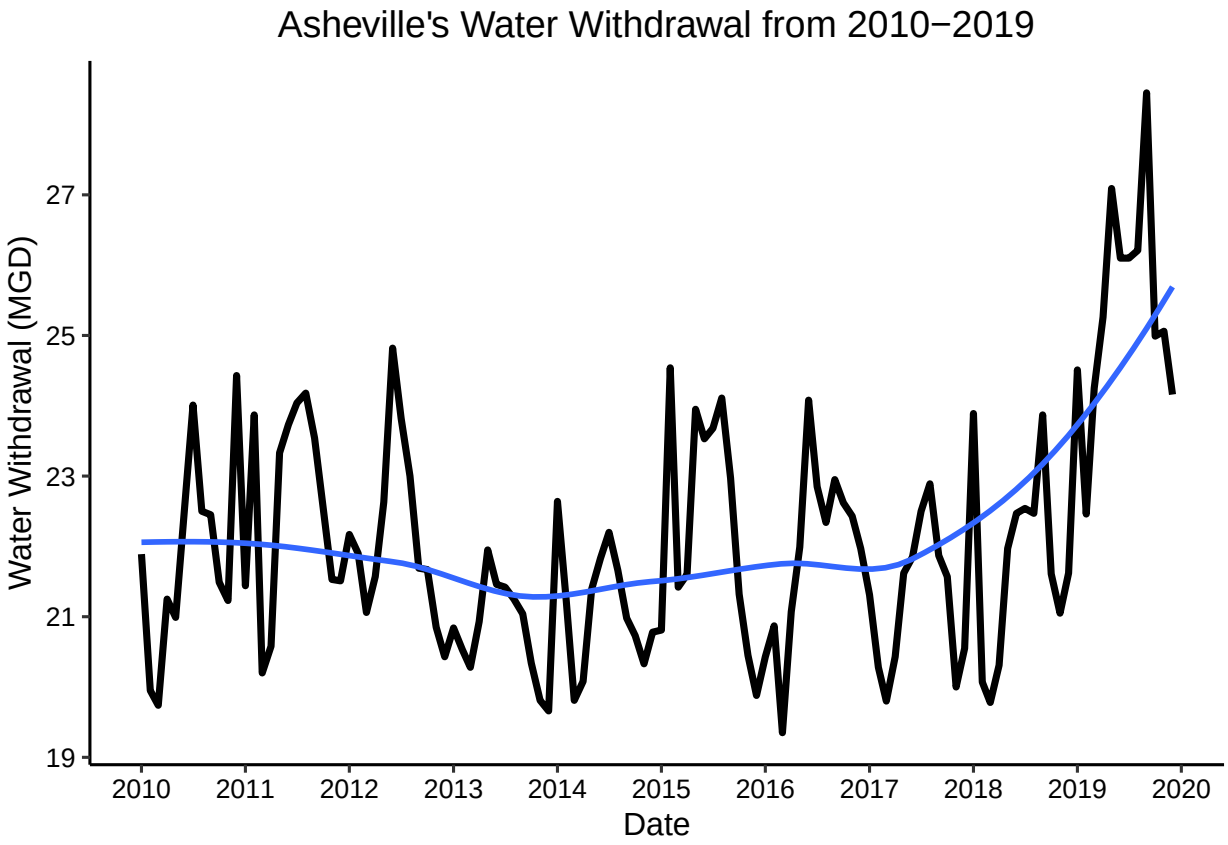
```
#9
#set inputs to scrape years 2010-2019 for Asheville site ("01-11-010")
the_years=rep(2010:2019)
PWSID = "01-11-010"

#using the Purrr's map function
Asheville_dfs <- map(the_years, scrape.it, the_pwsid=PWSID)

#conflate the returned dataframes into a single dataframe
Asheville_df <- bind_rows(Asheville_dfs)

#plot
ggplot(Asheville_df, aes(x=Date, y=Max.monthly.WD))+
  geom_line(size=1.25)+
  geom_smooth(method="loess",se=FALSE)+
  labs(title="Asheville's Water Withdrawal from 2010-2019",
       y="Water Withdrawal (MGD)")+
  scale_x_date(date_labels = "%Y", date_breaks = "1 year")
```

```
## `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? It appears that Asheville's water usage was relatively steady from 2010 to 2016 or 2017 but since then, it appears to be increasing fairly rapidly.