Assignment 09: Data Scraping

Elsie Liu

Total points:

OVERVIEW

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

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/Lsy/Box Sync/Duke/spring2022/872-EDA/GitRepository/Environmental_Data_Analytics_2022"
```

```
library(tidyverse)
library(lubridate)
library(rvest)
library(ggplot2)
library(plyr)
mytheme <- theme_classic() +
  theme(axis.text = element_text(color = "black"),
        legend.position = "bottom",plot.title = element_text(hjust = 0.5))
theme_set(mytheme)
```

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham's 2020 Municipal Local Water Supply Plan (LWSP):

- Navigate to https://www.ncwater.org/WUDC/app/LWSP/search.php
- Change the date from 2021 to 2020 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. (In other words, read the contents into an `rvest` webpage object.)

```
#2
DurhamURL <- read_html("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2020")
```

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:
- 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 <- DurhamURL %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()
pswid <- DurhamURL %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(5)") %>%
  html_text()
ownership <- DurhamURL %>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%
  html_text()
max.withdrawals.mgd <- DurhamURL %>%
  html_nodes("th~ td+ td") %>%
  html_text()
```

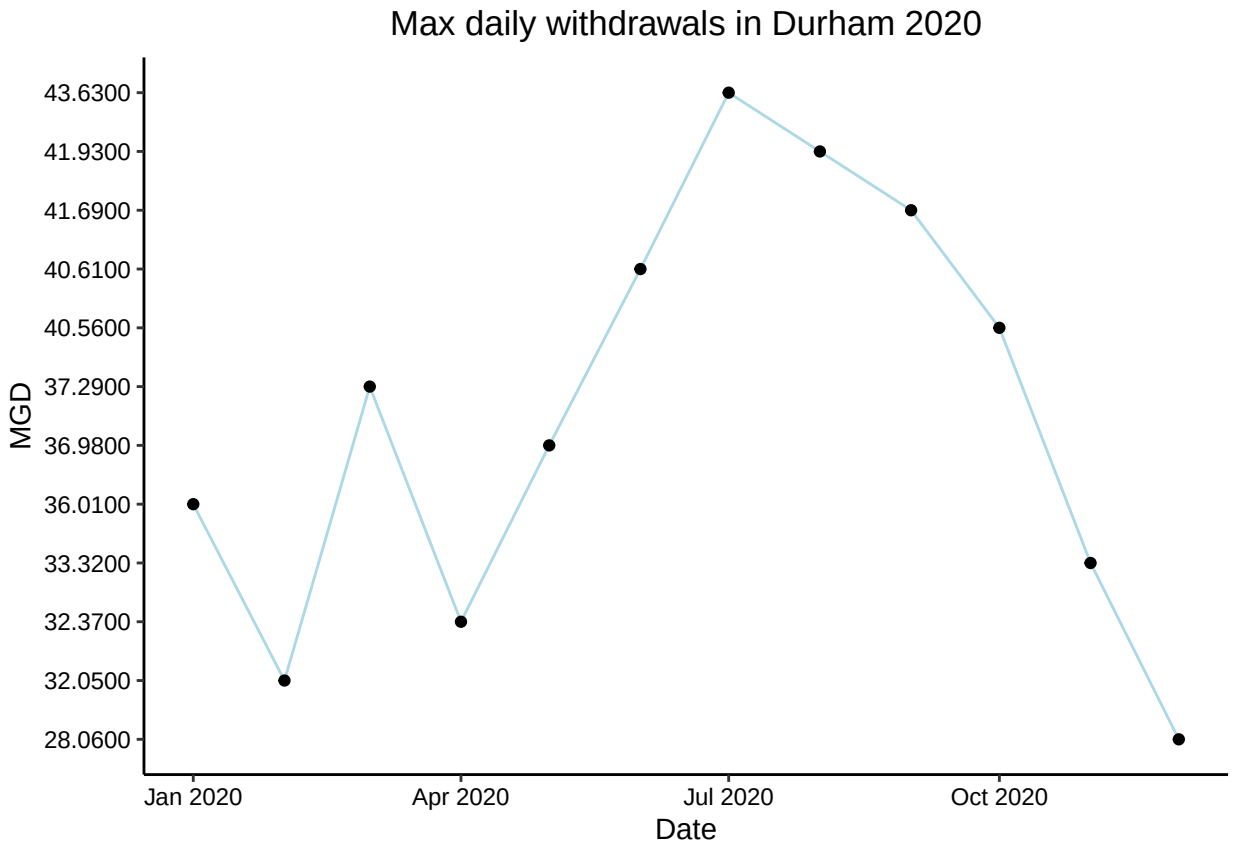
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   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
DWaterData <- data.frame("SystemName"=rep(water.system.name,12),"PSWID"=rep(pswid,12),"Ownership"=rep(ow
DWaterData$Date <- ym(with(DWaterData,paste(Year,Month,sep="-")))
#5
ggplot(data = DWaterData,aes(x = Date,y = MGD,group=1))+
  geom_line(color="lightblue")+
  geom_point()+
  ggtitle("Max daily withdrawals in Durham 2020")
```



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.

waterscraping <- function(siteID,Year){
  try(if(class(siteID)!="character") stop("Need siteID as string"))
  if (Year==2020) {
  URL <- read_html(paste("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=", siteID,sep = ""))
  }
  else {
  URL <- read_html(paste("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=", siteID, "&year=", Y
  }
  water.system.name <- URL %>%
    html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
    html_text()
  pswid <- URL %>%
    html_nodes("div+ table tr:nth-child(1) td:nth-child(5)") %>%
    html_text()
  ownership <- URL %>%
    html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%
    html_text()
  max.withdrawals.mgd <- URL %>%
    html_nodes("th~ td+ td") %>%
    html_text()
  WaterData <- data.frame("SystemName"=rep(water.system.name,12),"PSWID"=rep(pswid,12),"Ownership"=rep(
  WaterData$Date <- ym(with(WaterData,paste(Year,Month,sep="-")))
```

```
  return(WaterData)
}
```

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

```
#7
Durham15<- waterscraping(siteID = '03-32-010',2015)
Durham15
```

```
##    SystemName     PSWID    Ownership     MGD Month Year       Date
## 1      Durham 03-32-010 Municipality 40.2500     1 2015 2015-01-01
## 2      Durham 03-32-010 Municipality 53.1700     5 2015 2015-05-01
## 3      Durham 03-32-010 Municipality 40.0300     9 2015 2015-09-01
## 4      Durham 03-32-010 Municipality 43.5000     2 2015 2015-02-01
## 5      Durham 03-32-010 Municipality 57.0200     6 2015 2015-06-01
## 6      Durham 03-32-010 Municipality 38.7200    10 2015 2015-10-01
## 7      Durham 03-32-010 Municipality 43.1000     3 2015 2015-03-01
## 8      Durham 03-32-010 Municipality 41.6500     7 2015 2015-07-01
## 9      Durham 03-32-010 Municipality 43.5500    11 2015 2015-11-01
## 10     Durham 03-32-010 Municipality 49.6800     4 2015 2015-04-01
## 11     Durham 03-32-010 Municipality 44.7000     8 2015 2015-08-01
## 12     Durham 03-32-010 Municipality 48.7500    12 2015 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 the Asheville to Durham's water withdrawals.

```
#8
Asheville15<- waterscraping(siteID = '01-11-010',2015)
Asheville15
```

```
##    SystemName     PSWID    Ownership     MGD Month Year       Date
## 1   Asheville 01-11-010 Municipality 20.8100     1 2015 2015-01-01
## 2   Asheville 01-11-010 Municipality 23.9500     5 2015 2015-05-01
## 3   Asheville 01-11-010 Municipality 22.9700     9 2015 2015-09-01
## 4   Asheville 01-11-010 Municipality 24.5400     2 2015 2015-02-01
## 5   Asheville 01-11-010 Municipality 23.5300     6 2015 2015-06-01
## 6   Asheville 01-11-010 Municipality 21.3200    10 2015 2015-10-01
## 7   Asheville 01-11-010 Municipality 21.4200     3 2015 2015-03-01
## 8   Asheville 01-11-010 Municipality 23.6800     7 2015 2015-07-01
## 9   Asheville 01-11-010 Municipality 20.4500    11 2015 2015-11-01
## 10  Asheville 01-11-010 Municipality 21.6000     4 2015 2015-04-01
## 11  Asheville 01-11-010 Municipality 24.1100     8 2015 2015-08-01
## 12  Asheville 01-11-010 Municipality 19.8800    12 2015 2015-12-01
```

```
DurAsh15 <- full_join(Durham15,Asheville15)
```

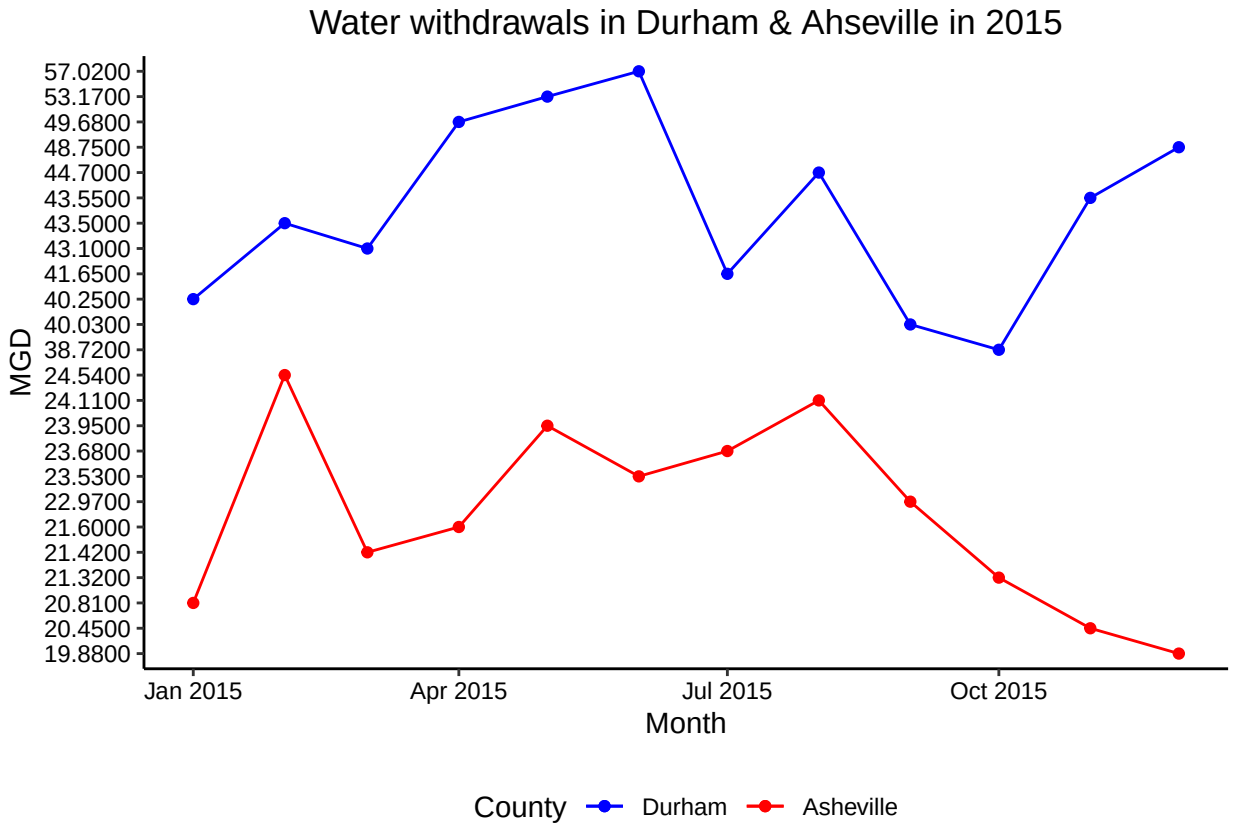
```
## Joining, by = c("SystemName", "PSWID", "Ownership", "MGD", "Month", "Year", "Date")
```

```
DurAsh15 <- reshape(DurAsh15, idvar = "Month", v.names = "MGD", timevar = "SystemName", direction = "wi
```

```
## Warning in reshapeWide(data, idvar = idvar, timevar = timevar, varying =
## varying, : some constant variables (PSWID) are really varying
```

```
ggplot(data = DurAsh15)+
  geom_point(aes(x = Date,y = MGD.Durham,color="blue"))+
  geom_line(aes(x = Date,y = MGD.Durham,color="blue",group=1))+
```

```
  geom_point(aes(x = Date,y = MGD.Asheville,color="red"))+
  geom_line(aes(x = Date,y = MGD.Asheville,color="red",group=1))+
  labs(title = "Water withdrawals in Durham & Ahseville in 2015", x = "Month", y = "MGD", color = "Coun
  scale_color_manual(labels = c("Durham", "Asheville"),values = c("blue","red"))
```

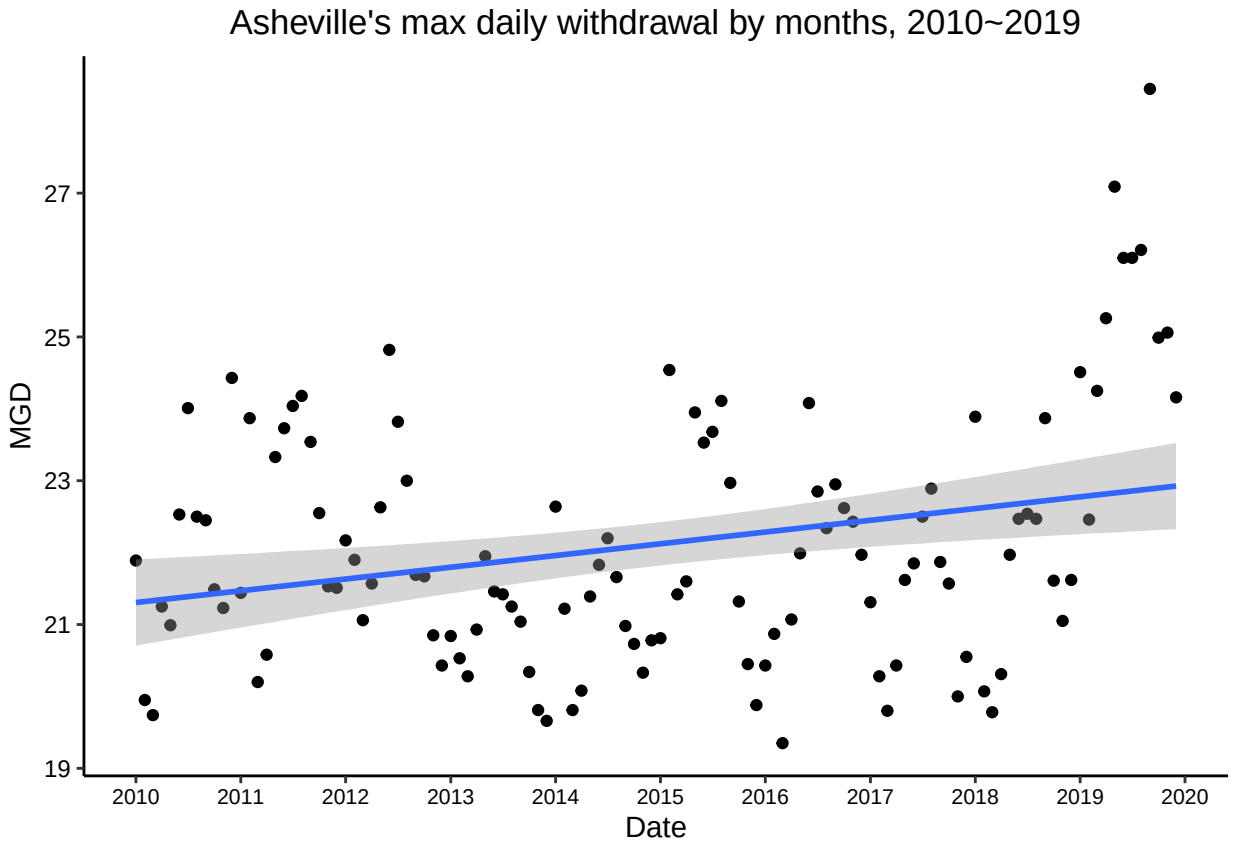


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

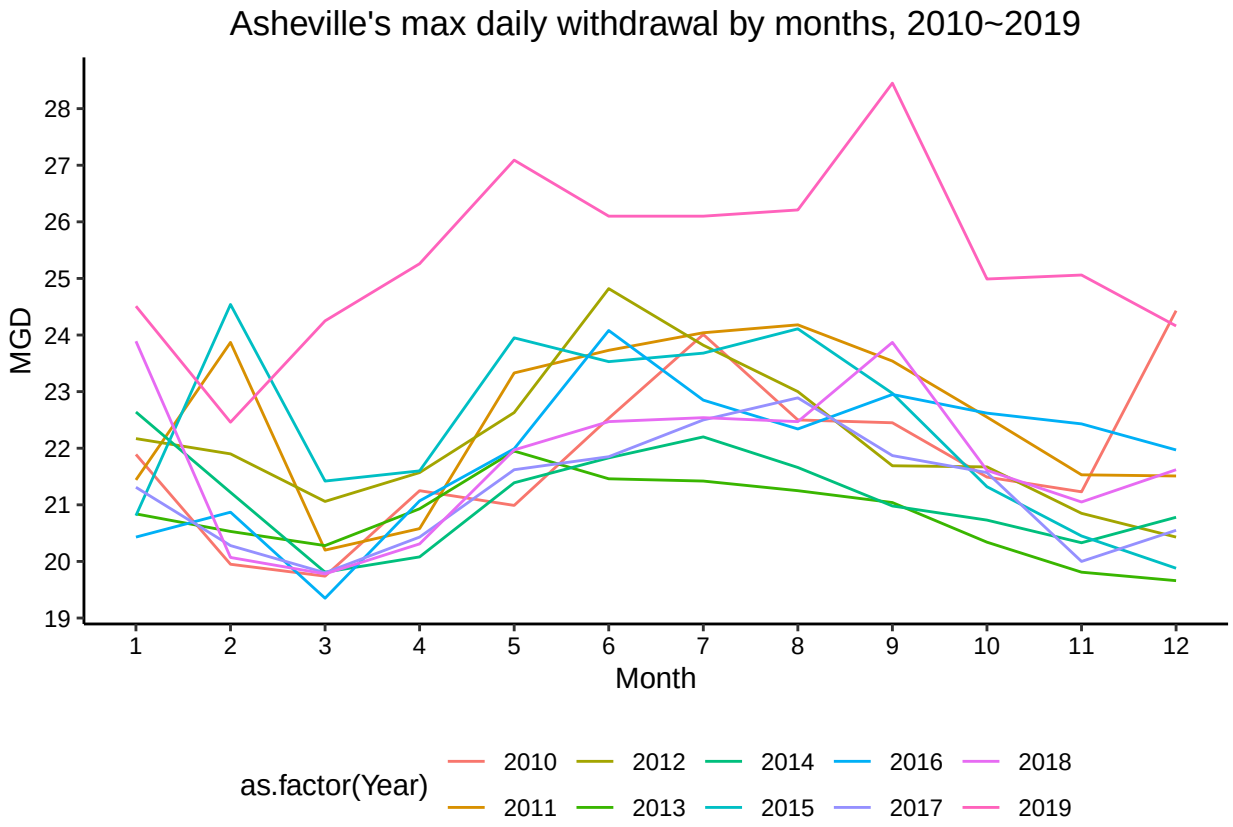
```
#9
Ash10 <- waterscraping('01-11-010',2010)
Ash11 <- waterscraping('01-11-010',2011)
Ash12 <- waterscraping('01-11-010',2012)
Ash13 <- waterscraping('01-11-010',2013)
Ash14 <- waterscraping('01-11-010',2014)
Ash15 <- waterscraping('01-11-010',2015)
Ash16 <- waterscraping('01-11-010',2016)
Ash17 <- waterscraping('01-11-010',2017)
Ash18 <- waterscraping('01-11-010',2018)
Ash19 <- waterscraping('01-11-010',2019)
Ash1019 <- rbind(Ash10,Ash11,Ash12,Ash13,Ash14,Ash15,Ash16,Ash17,Ash18,Ash19)
Ash1019$MGD <- as.numeric(Ash1019$MGD)

par(mfrow=c(2,2))
ggplot(data = Ash1019,aes(x = Date, y = MGD))+
  geom_point()+
  geom_smooth(method = lm)+
  scale_x_date(date_labels = "%Y",date_breaks = "1 year")+
  labs(title = "Asheville's max daily withdrawal by months, 2010~2019")+
```

```
  theme(axis.text.x = element_text(size = 8))
```

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




```
ggplot(data = Ash1019)+
  geom_line(aes(x = Month, y = MGD, group=Year,color=as.factor(Year)))+
  scale_y_continuous(breaks = seq(19, 29, by = 1))+
  scale_x_continuous(breaks = seq(1, 12, by = 1))+
  labs(title = "Asheville's max daily withdrawal by months, 2010~2019")
```



Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time?

Answer: Asheville has a increasing trend in water usage over years and also has a seasonal trend within each year.