

Assignment 09: Data Scraping

Langston Alexander

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] "C:/Users/lwa8/Documents/R/ENV872/Environmental_Data_Analytics_2022/Assignments"

library(tidyverse)
library(rvest)

## Warning: package 'rvest' was built under R version 4.1.3

library(lubridate)
library(purrr)

mytheme <- theme_classic() +
  theme(axis.text = element_text(color = "black"))
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 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')
```

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()  
pswid <- 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()
```

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 withdrawal 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 <- 2020  
maxuse.df <- data.frame(  
  "Month" = as.factor(month),  
  "Max-Withdrawal" = as.numeric(max.withdrawals.mgd),  
  "System_Name" = water.system.name,  
  "Ownership" = ownership,
```

```

"PSWID" = pswid,
'Date' = my(paste(month,"-",year)))

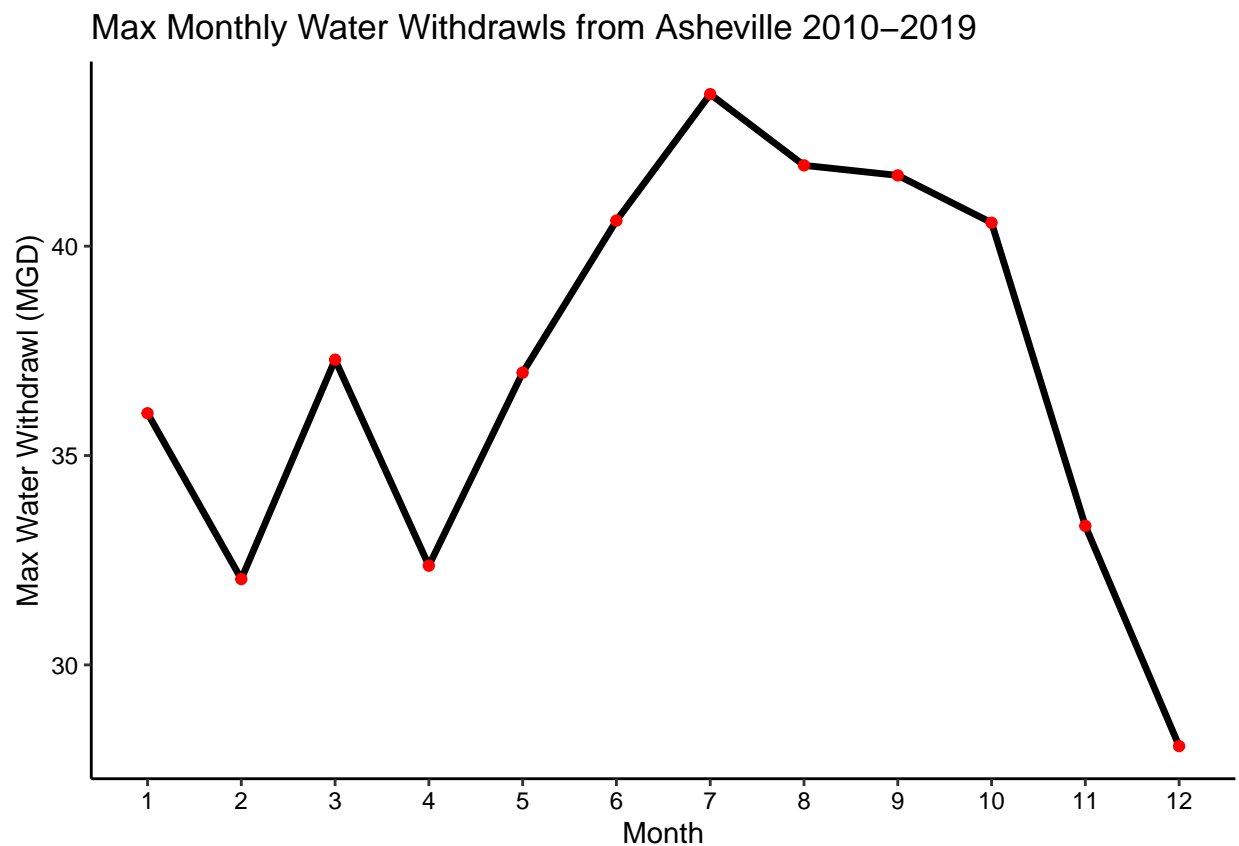
maxuse.df <- arrange(maxuse.df, Month)

maxuse.df <- maxuse.df %>%
  mutate(
    MonthName = c(
      "Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sept", "Oct", "Nov", "Dec"
    )
  )

#5

ggplot(maxuse.df, aes(x = Month, y = Max_Withdrawl, group = 1))+
  geom_line(size = 1.2)+
  geom_point(color = "red")+
  labs(title = "Max Monthly Water Withdrawls from Asheville 2010-2019",
       y = "Max Water Withdrawl (MGD)")

```



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.

maxuse.scrape <- function(PSWID, Year){

```

```

#Fetch Website

webpage <- read_html(paste0(
  'https://www.ncwater.org/WUDC/app/LWSP/report.php?pswid=',
  PSWID, '&year=',
  Year))

#Scrape element data

water.system.name <- webpage %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()
pswid <- 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()

#Convert to df

month <- c(1,5,9,2,6,10,3,7,11,4,8,12)
maxuse <- data.frame(
  "Month" = as.factor(month),
  "Max-Withdrawl" = as.numeric(max.withdrawals.mgd),
  "System_Name" = water.system.name,
  "Ownership" = ownership,
  "PSWID" = pswid,
  'Date' = my(paste(month,"-",Year)))

maxuse <- arrange(maxuse, Month)

return(maxuse)
}

maxuse.scrape('03-32-010', 2019)

```

##	Month	Max-Withdrawl	System_Name	Ownership	PSWID	Date
## 1	1	29.62	Durham Municipality	03-32-010	2019-01-01	
## 2	2	32.39	Durham Municipality	03-32-010	2019-02-01	
## 3	3	36.43	Durham Municipality	03-32-010	2019-03-01	
## 4	4	32.60	Durham Municipality	03-32-010	2019-04-01	
## 5	5	35.73	Durham Municipality	03-32-010	2019-05-01	
## 6	6	37.86	Durham Municipality	03-32-010	2019-06-01	
## 7	7	46.02	Durham Municipality	03-32-010	2019-07-01	
## 8	8	42.05	Durham Municipality	03-32-010	2019-08-01	
## 9	9	54.07	Durham Municipality	03-32-010	2019-09-01	

```
## 10    10      44.35    Durham Municipality 03-32-010 2019-10-01
## 11    11      36.06    Durham Municipality 03-32-010 2019-11-01
## 12    12      31.20    Durham Municipality 03-32-010 2019-12-01
```

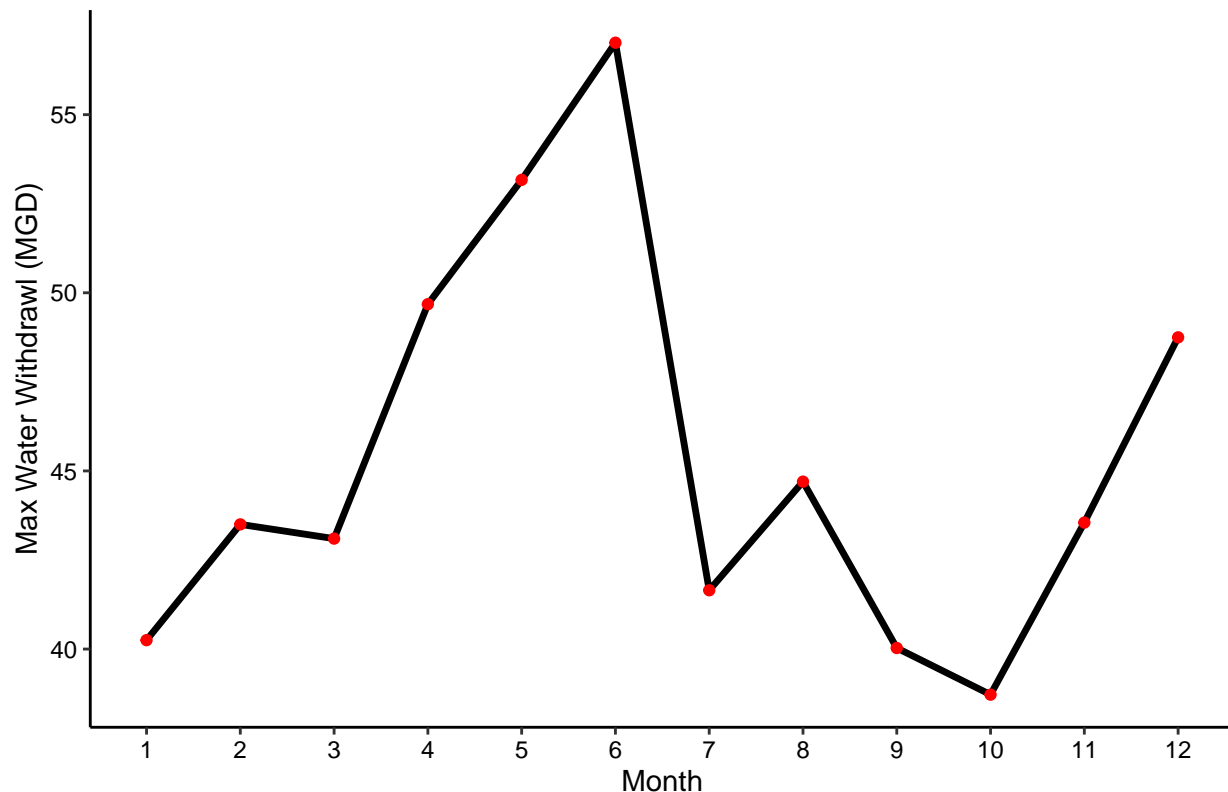
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID='03-32-010') for each month in 2015

#7

```
maxuseDurham.df <- maxuse.scrape("03-32-010", 2015)

ggplot(maxuseDurham.df, aes(x = Month, y=Max-Withdrawl, group = 1))+
  geom_line(size = 1.2)+
  geom_point(color = "red")+
  labs(title = "Max Monthly Water Withdrawls from Durham in 2015",
       y = "Max Water Withdrawl (MGD)")
```

Max Monthly Water Withdrawls from Durham in 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

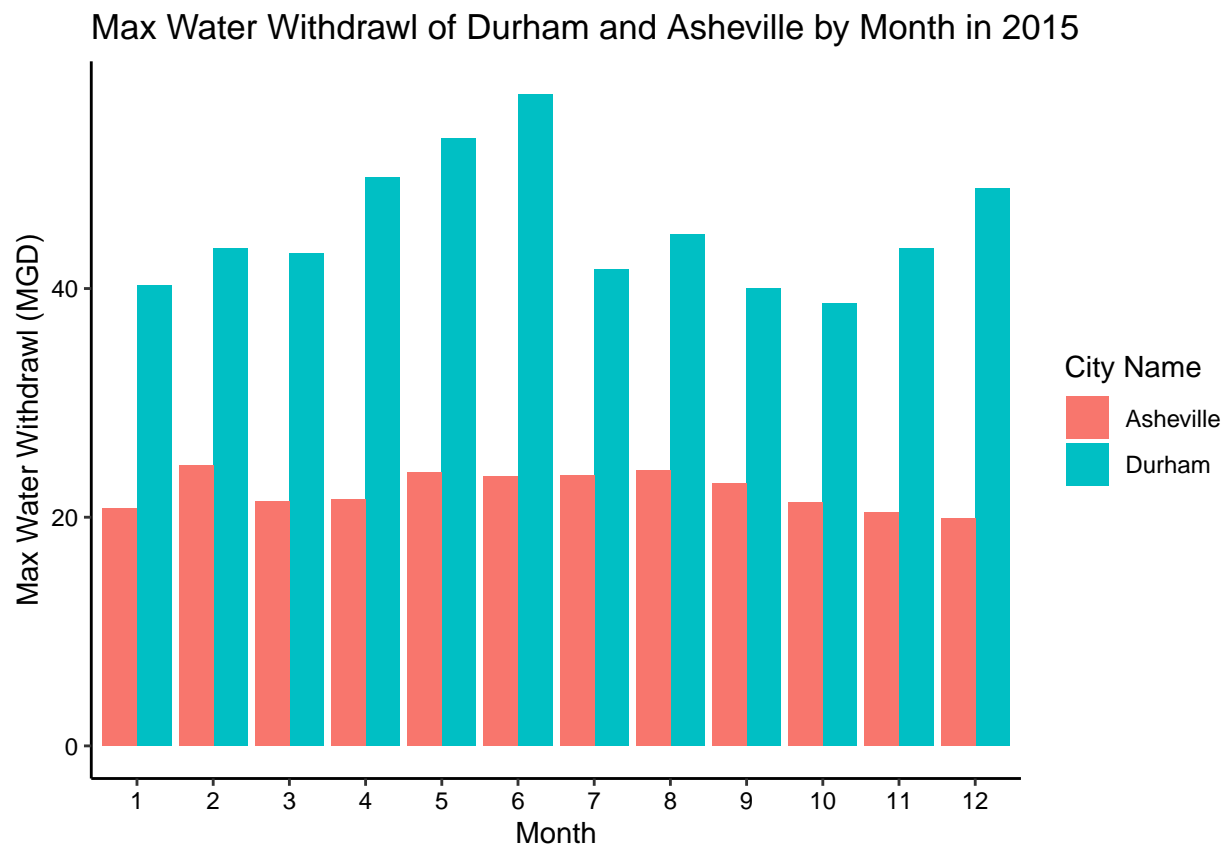
```
maxuseAshville.df <- maxuse.scrape("01-11-010", 2015)
maxuseAshville.df
```

```
##      Month Max-Withdrawl System_Name  Ownership   PSWID      Date
## 1      1      20.81    Asheville Municipality 01-11-010 2015-01-01
## 2      2      24.54    Asheville Municipality 01-11-010 2015-02-01
## 3      3      21.42    Asheville Municipality 01-11-010 2015-03-01
```

```
## 4      4      21.60 Asheville Municipality 01-11-010 2015-04-01
## 5      5      23.95 Asheville Municipality 01-11-010 2015-05-01
## 6      6      23.53 Asheville Municipality 01-11-010 2015-06-01
## 7      7      23.68 Asheville Municipality 01-11-010 2015-07-01
## 8      8      24.11 Asheville Municipality 01-11-010 2015-08-01
## 9      9      22.97 Asheville Municipality 01-11-010 2015-09-01
## 10     10      21.32 Asheville Municipality 01-11-010 2015-10-01
## 11     11      20.45 Asheville Municipality 01-11-010 2015-11-01
## 12     12      19.88 Asheville Municipality 01-11-010 2015-12-01
```

```
DurAsheMaxUSE.df <- bind_rows(maxuseDurham.df, maxuseAshville.df)
```

```
ggplot(DurAsheMaxUSE.df, aes(x = Month, y = Max-Withdrawl, fill = System_Name))+
  geom_bar(position = 'dodge', stat = "identity")+
  labs(
    title = "Max Water Withdrawl of Durham and Asheville by Month in 2015",
    y = "Max Water Withdrawl (MGD)",
    fill = "City Name")
```



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

```
the_years <- rep(2010:2019)
the_pswid <- "01-11-010"
```

```

ashw_dfs <- map(the_years, maxuse.scrape, PSWID = the_pswid)

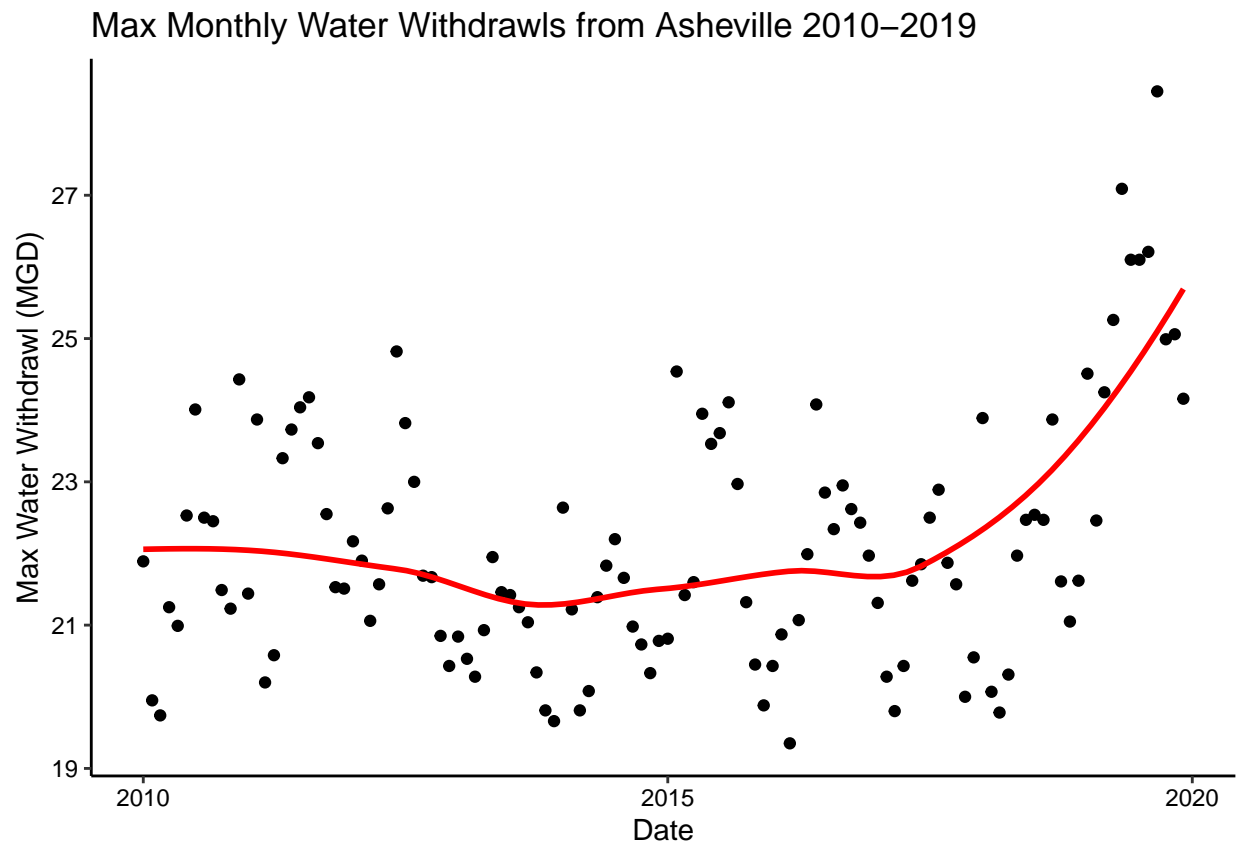
ashe_df <- bind_rows(ashw_dfs)

ashe_df$year <- as.factor(year(ydm(ashe_df$Date)))

ggplot(ashe_df, aes(x = Date, y = Max_Withdrawl))+
  geom_point()+
  geom_smooth(color = "red", se = F)+
  labs(title = "Max Monthly Water Withdrawls from Asheville 2010-2019",
       y = "Max Water Withdrawl (MGD)")

```

```
## `geom_smooth()` using method = 'loess' and 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 looks like there wasn't a clear trend in water usage from 2010 until 2017. In 2017 water usage started to rise and continued to do so through 2019.