Assignment 10: Data Scraping

Sokna Kry

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(rvest)
library(dbplyr)
library(lubridate)
library(ggplot2)

getwd()
```

```
## [1] "/Users/sokna/Documents/EDA-Spring2023"
```

```
# Set theme
mytheme <- theme_classic() +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
theme_set(mytheme)
```

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

```
the_registrant <- webpage %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()
the_registrant
```

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

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), with the first value being "36.1000".

```
#3
water.system.name <- webpage %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()
water.system.name
```

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

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

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

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

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

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

```
##  [1] "36.1000" "43.4200" "52.4900" "30.5000" "42.5900" "34.8800" "39.9100"
##  [8] "43.3200" "32.5300" "34.6600" "41.8000" "37.5300"
```

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.

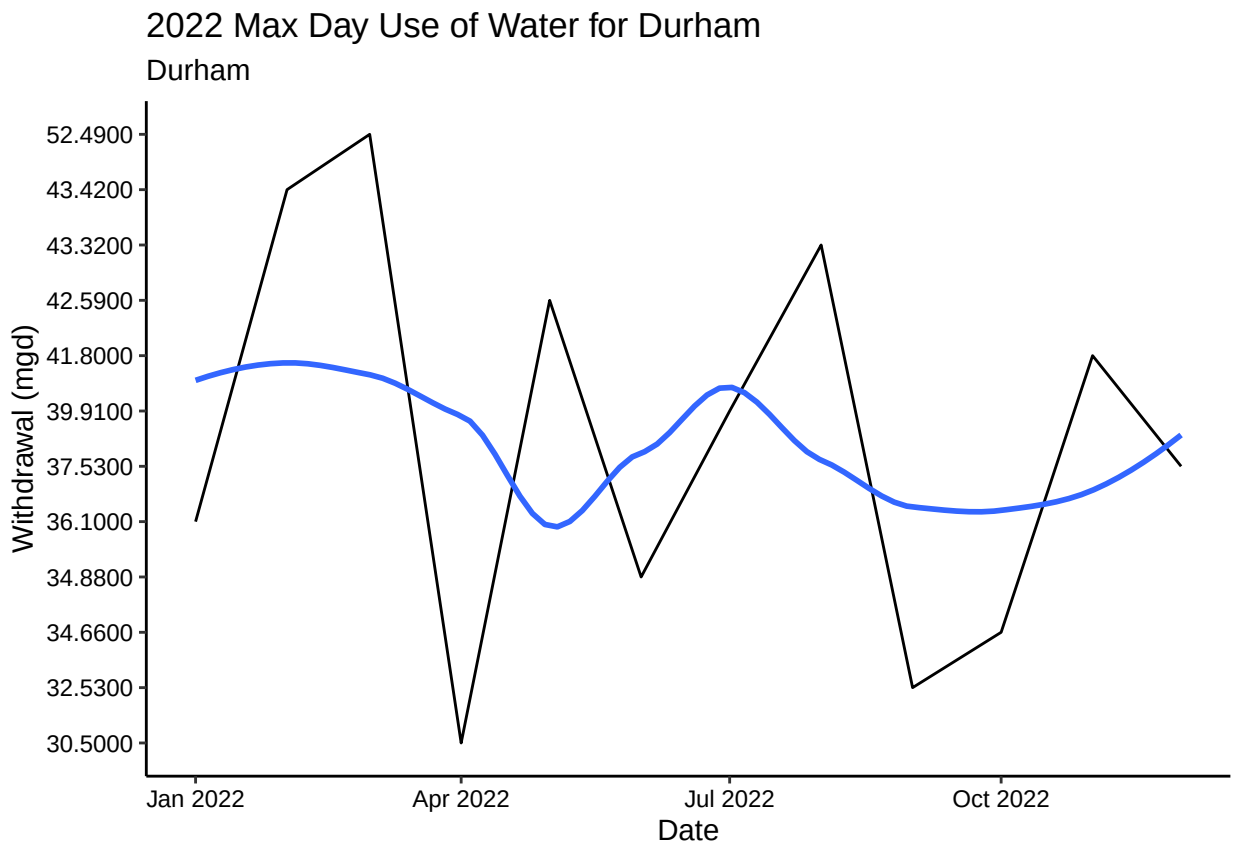
   NOTE: It's likely you won't be able to scrape the monthly widthrawal data in chronological order. You can overcome this by creating a month column manually assigning values in the order the data are scraped: "Jan", "May", "Sept", "Feb", etc... Or, you could scrape month values from the web page...

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

```
#4
df_max.withdrawals.mgd <- data.frame("Month" = rep(1:12),
                             "Year" = rep(2022,12),
                             "max.withdrawals_mgd" = as.numeric(max.withdrawals.mgd))

df_max.withdrawals.mgd <- data.frame("Month" = rep(1:12),
                             "Year" = rep(2022,12),
                             "max.withdrawals_mgd" = as.numeric(max.withdrawals.mgd))

df_max.withdrawals.mgd <- df_max.withdrawals.mgd %>%
  mutate(water.system.name = !!water.system.name,
         PWSID = !!PWSID,
         ownership = !!ownership,
         Date = my(paste(Month,"-",Year)))
```

```
#5

#Plot
ggplot(df_max.withdrawals.mgd,aes(x=Date,y=max.withdrawals.mgd, group=water.system.name)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2022 Max Day Use of Water for",water.system.name),
       subtitle = water.system.name,
       y="Withdrawal (mgd)",
       x="Date")
```

```
## `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.

the_base_url <- 'https://www.ncwater.org/WUDC/app/LWSP/report.php?'
PWSID <- '03-32-010'
the_year <- 2015
the_scrape_url <- paste0(the_base_url, "pwsid=",PWSID, "&year=",the_year)
print(the_scrape_url)
```

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

```
scrape.it <- function(the_year, water.system.name)

  #Retrieve the website contents
website <- read_html(paste0(the_base_url, "pwsid=",PWSID, "&year=",the_year))
website<- read_html(the_scrape_url)


  #Scrape the data items
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.withdrawals <- website %>% html_nodes("th~ td+ td") %>% html_text()

  #Convert to a dataframe
df_withdrawals <- data.frame("Month" = rep(1:12),
                             "Year" = rep(the_year,12),
                             "max.withdrawals" = as.numeric(max.withdrawals)) %>% mutate(water.system
        PWSID = !!PWSID,
        ownership = !!ownership,
        Date = my(paste(Month,"-",Year)))

  #Return the dataframe
return(df_withdrawals)
```

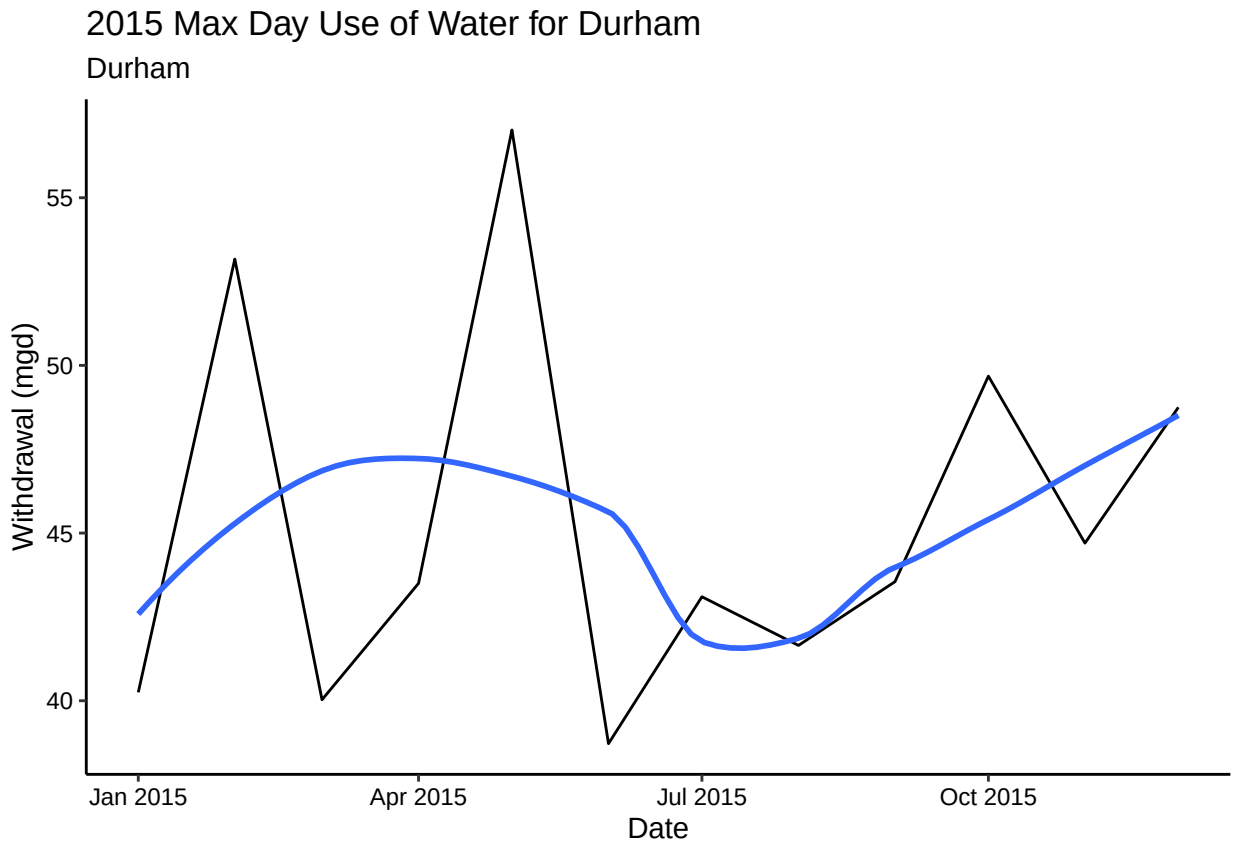
```
##    Month Year max.withdrawals water.system.name     PWSID    ownership
## 1      1 2015           40.25            Durham 03-32-010 Municipality
## 2      2 2015           53.17            Durham 03-32-010 Municipality
## 3      3 2015           40.03            Durham 03-32-010 Municipality
## 4      4 2015           43.50            Durham 03-32-010 Municipality
## 5      5 2015           57.02            Durham 03-32-010 Municipality
## 6      6 2015           38.72            Durham 03-32-010 Municipality
## 7      7 2015           43.10            Durham 03-32-010 Municipality
## 8      8 2015           41.65            Durham 03-32-010 Municipality
## 9      9 2015           43.55            Durham 03-32-010 Municipality
## 10    10 2015           49.68            Durham 03-32-010 Municipality
## 11    11 2015           44.70            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
```

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

```
#7

ggplot(df_withdrawals,aes(x=Date,y=max.withdrawals, group=water.system.name)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2015 Max Day Use of Water for",water.system.name),
       subtitle = water.system.name,
       y="Withdrawal (mgd)",
       x="Date")
```

```
## ‘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 Asheville’s to Durham’s water withdrawals.

```
#8

the_base_url <- 'https://www.ncwater.org/WUDC/app/LWSP/report.php?'
```

```
PWSID <- '01-11-010'
the_year <- 2015
the_scrape_url2 <- paste0(the_base_url, "pwsid=",PWSID, "&year=",the_year)
print(the_scrape_url2)
```

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

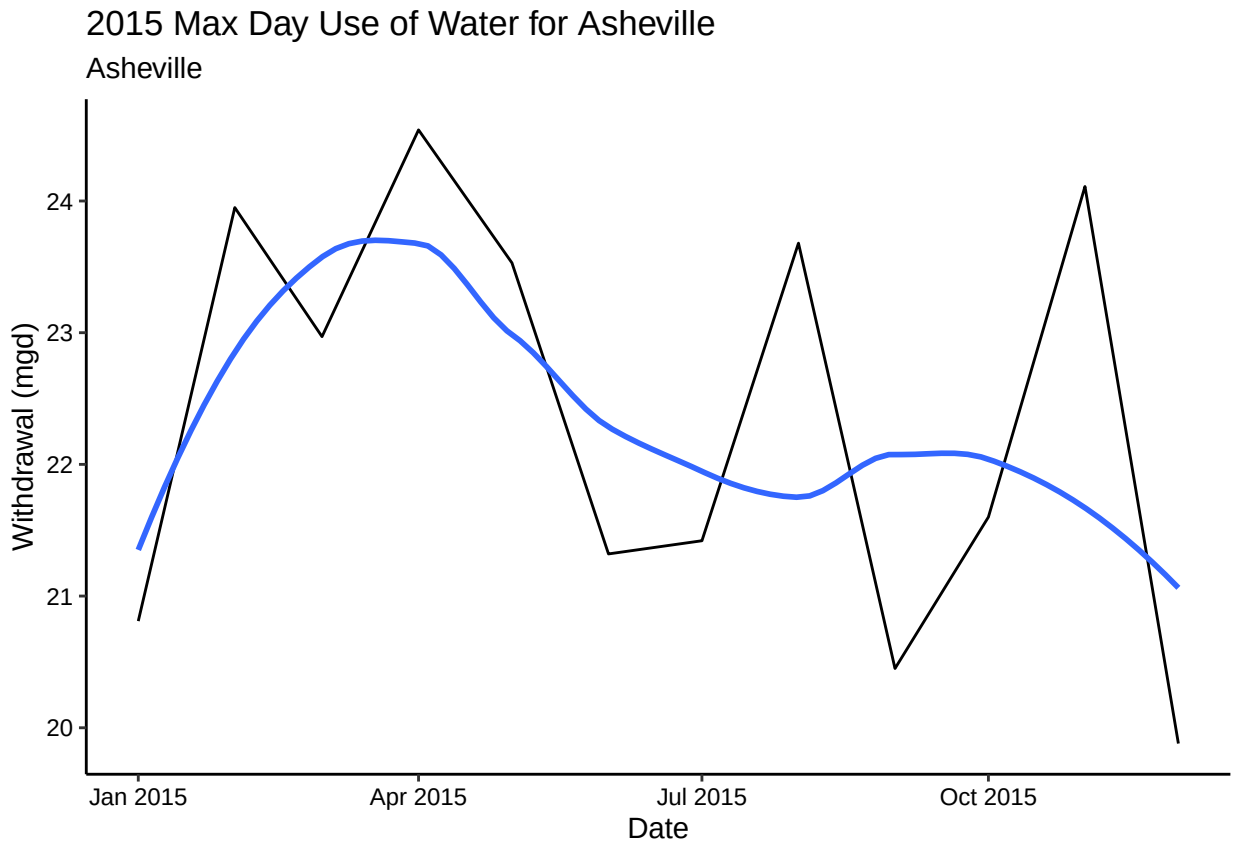
```
scrape.it <- function(the_year, water.system.name)

  #Retrieve the website contents
  website <- read_html(paste0(the_base_url, "pwsid=",PWSID, "&year=",the_year))
  website<- read_html(the_scrape_url2)


  #Scrape the data items
  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.withdrawals <- website %>% html_nodes("th~ td+ td") %>% html_text()

  #Convert to a dataframe
  df_withdrawals <- data.frame("Month" = rep(1:12),
                               "Year" = rep(the_year,12),
                               "max.withdrawals" = as.numeric(max.withdrawals)) %>% mutate(water.system
        PWSID = !!PWSID,
        ownership = !!ownership,
        Date = my(paste(Month,"-",Year)))


  #Return the dataframe
  return(df_withdrawals)
```

```
##    Month Year max.withdrawals water.system.name     PWSID    ownership
## 1      1 2015           20.81         Asheville 01-11-010 Municipality
## 2      2 2015           23.95         Asheville 01-11-010 Municipality
## 3      3 2015           22.97         Asheville 01-11-010 Municipality
## 4      4 2015           24.54         Asheville 01-11-010 Municipality
## 5      5 2015           23.53         Asheville 01-11-010 Municipality
## 6      6 2015           21.32         Asheville 01-11-010 Municipality
## 7      7 2015           21.42         Asheville 01-11-010 Municipality
## 8      8 2015           23.68         Asheville 01-11-010 Municipality
## 9      9 2015           20.45         Asheville 01-11-010 Municipality
## 10    10 2015           21.60         Asheville 01-11-010 Municipality
## 11    11 2015           24.11         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
```

```
ggplot(df_withdrawals,aes(x=Date,y=max.withdrawals, group=water.system.name)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2015 Max Day Use of Water for",water.system.name),
       subtitle = water.system.name,
       y="Withdrawal (mgd)",
       x="Date")
```

```
## `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 2021.Add a smoothed line to the plot (method = 'loess').

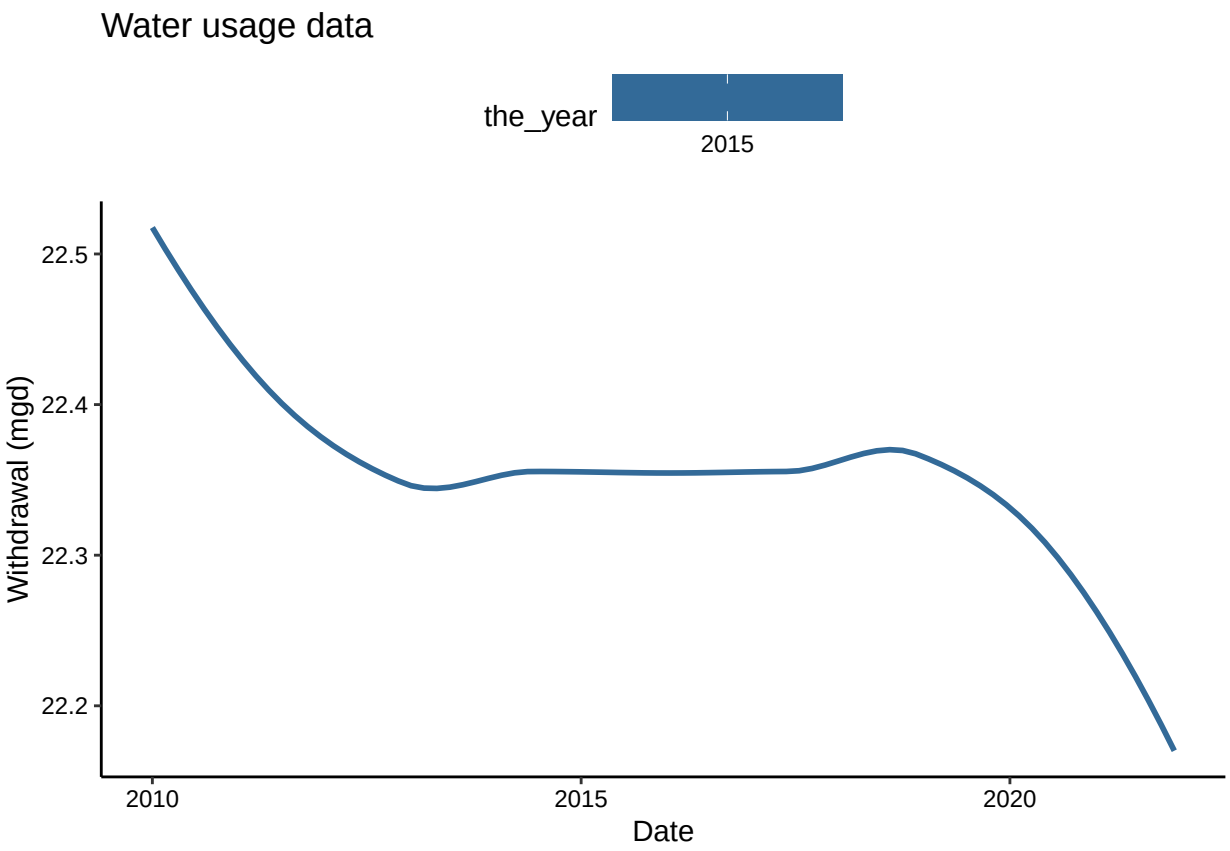
   TIP: See Section 3.2 in the "09_Data_Scraping.Rmd" where we apply "map2()" to iteratively run a function over two inputs. Pipe the output of the map2() function to `bindrows()` to combine the dataframes into a single one.

```r
#9

#Code & function from Q#8
the_base_url <- 'https://www.ncwater.org/WUDC/app/LWSP/report.php?'
PWSID <- '01-11-010'
the_year <- 2015
the_scrape_url2 <- paste0(the_base_url, "pwsid=",PWSID, "&year=",the_year)
print(the_scrape_url2)
```

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

```r
scrape.it <- function(the_year, water.system.name){

  #Retrieve the website contents
  website <- read_html(paste0(the_base_url, "pwsid=",PWSID, "&year=",the_year))
  website<- read_html(the_scrape_url2)


  #Scrape the data items
  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.withdrawals <- website %>% html_nodes("th~ td+ td") %>% html_text()

  #Convert to a dataframe
  df_withdrawals <- data.frame("Month" = rep(1:12),
                               "Year" = rep(the_year,12),
                               "max.withdrawals" = as.numeric(max.withdrawals)) %>% mutate(water.system
        PWSID = !!PWSID,
        ownership = !!ownership,
        Date = my(paste(Month,"-",Year)))

  #Return the dataframe
  return(df_withdrawals)
}

#Create dataframe

df_years<- seq(2010,2021) %>%
  map(scrape.it) %>%
  bind_rows()

ggplot(df_years,aes(x=Date,y=max.withdrawals,color=the_year)) +
  #geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = "Water usage data",
       y="Withdrawal (mgd)",
       x="Date")
```

```
## `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? Yes. Asheville has a trend in water usage over time. The water usage is decreasing over time.