Assignment 10: Data Scraping

Brian Mulu Mutua

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 Loading the packages
library(tidyverse)
library(lubridate)
library(here)
library(rvest)

#Checking working directory
getwd()
```

```
## [1] "C:/Users/bmm100/Documents/EDE_Fall2023"
```

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 Reading the website contents into an rvest webpage object
ncLWSP_website <-
↪  read_html('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2022')

#Displaying object contents to demonstrate successful read
glimpse(ncLWSP_website)
```

```
## List of 2
##  $ node:<externalptr>
##  $ doc :<externalptr>
##  - attr(*, "class")= chr [1:2] "xml_document" "xml_node"
```

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 Scraping the required data values and storing as text
the_water_system_name <- ncLWSP_website %>%
  html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>%
  html_text()
the_pwsid <- ncLWSP_website %>%
  html_nodes('td tr:nth-child(1) td:nth-child(5)') %>%
  html_text()
the_ownership <- ncLWSP_website %>%
  html_nodes('div+ table tr:nth-child(2) td:nth-child(4)') %>%
  html_text()
the_max_day_use <- ncLWSP_website %>%
  html_nodes('th~ td+ td') %>%
  html_text()

#Displaying scraped values
the_water_system_name
```

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

```
the_pwsid
```

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

```
the_ownership
```

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

```
glimpse(the_max_day_use)
```

```
##  chr [1:12] "36.1000" "43.4200" "52.4900" "30.5000" "42.5900" "34.8800" ...
```

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.

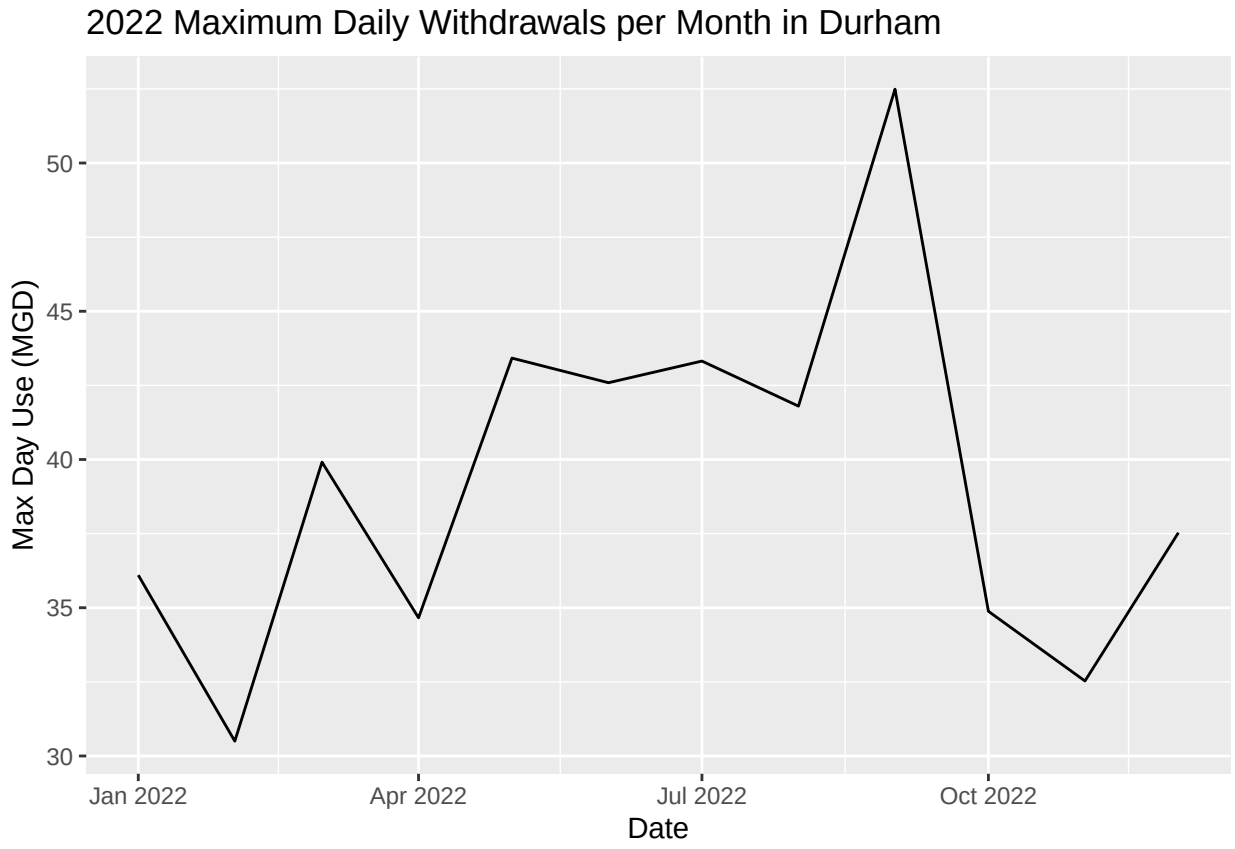
   NOTE: It's likely you won't be able to scrape the monthly widthrawal data in chronological order. You can overcome this by creating a month column manually assigning values in the order the data are scraped: "Jan", "May", "Sept", "Feb", etc... Or, you could scrape month values from the web page...

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

```
#4 Creating a dataframe from scraped data
df_maxdailyuse <- data.frame("Month" = c(1,5,9,2,6,10,3,7,11,4,8,12),
                             "Year" = rep(2022,12),
                             "Maximum_Day_Use_MGD" = as.numeric(the_max_day_use)) %>%
  mutate(Water_System_Name = !!the_water_system_name,
         PWSID = !!the_pwsid,
         Ownership = !!the_ownership,
         Date = my(paste(Month,"-",Year)))

#Displaying contents of created dataframe
glimpse(df_maxdailyuse)
```

```
## Rows: 12
## Columns: 7
## $ Month               <dbl> 1, 5, 9, 2, 6, 10, 3, 7, 11, 4, 8, 12
## $ Year                <dbl> 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 20~
## $ Maximum_Day_Use_MGD <dbl> 36.10, 43.42, 52.49, 30.50, 42.59, 34.88, 39.91, 4~
## $ Water_System_Name   <chr> "Durham", "Durham", "Durham", "Durham", "Durham", ~
## $ PWSID               <chr> "03-32-010", "03-32-010", "03-32-010", "03-32-010"~
## $ Ownership           <chr> "Municipality", "Municipality", "Municipality", "M~
## $ Date                <date> 2022-01-01, 2022-05-01, 2022-09-01, 2022-02-01, 20~
```

```r
#5 Creating line plot of the maximum daily withdrawals across the months for 2022
maxdailywithdrawals.plot <- df_maxdailyuse %>%
  ggplot(aes(x = Date, y = Maximum_Day_Use_MGD)) +
  geom_line() +
  labs(title = '2022 Maximum Daily Withdrawals per Month in Durham',
       x = 'Date',
       y = 'Max Day Use (MGD)')

maxdailywithdrawals.plot
```



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**.

```r
#6. Constructing PWSID scraping function

scrape.pwsid <- function(selected_pwsid, selected_year){

  #Retrieving website contents
  ncLWSP_website <-
↪  read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',
                                  selected_pwsid, '&year=', selected_year))

  #Scraping the data items using the element address variables determined from Question 3
```

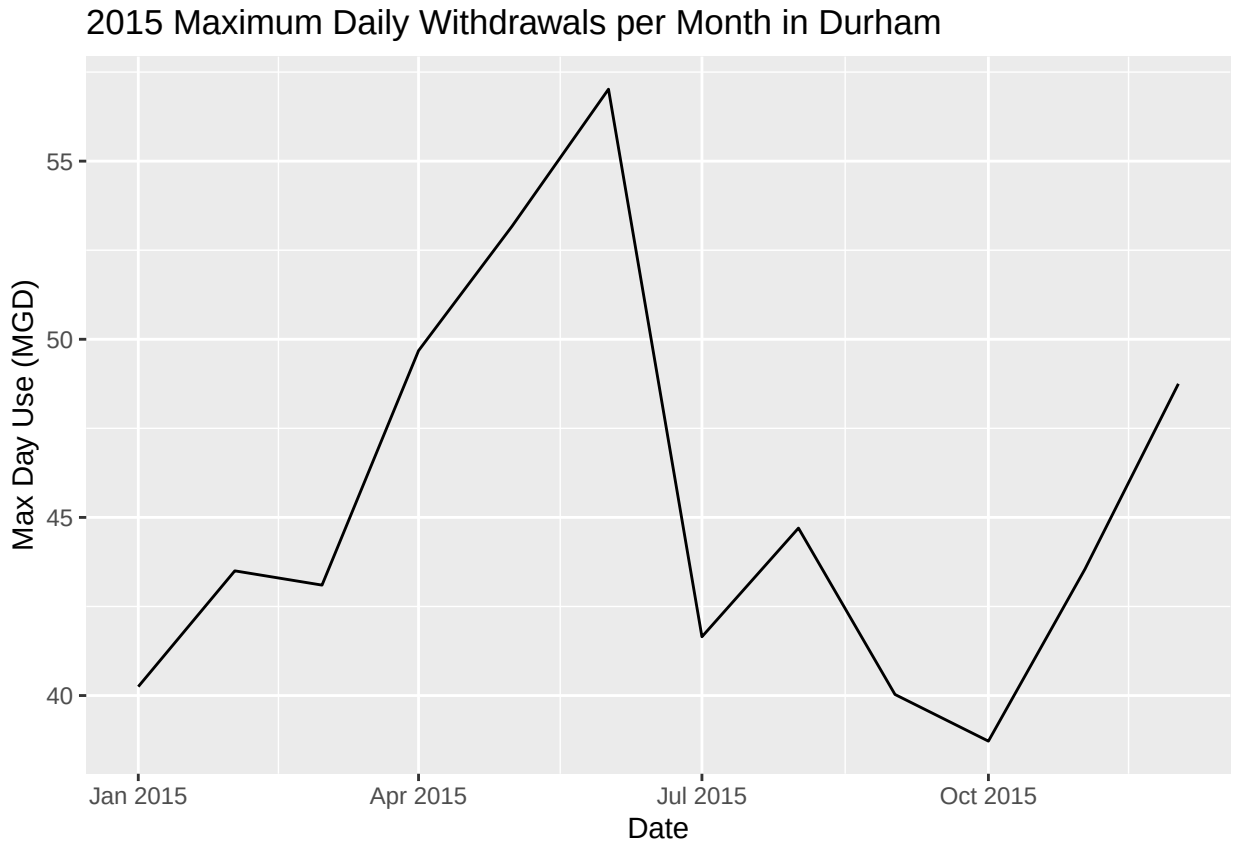
```
  the_water_system_name <- ncLWSP_website %>%
    html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>%
    html_text()
  the_pwsid <- ncLWSP_website %>%
    html_nodes('td tr:nth-child(1) td:nth-child(5)') %>%
    html_text()
  the_ownership <- ncLWSP_website %>%
    html_nodes('div+ table tr:nth-child(2) td:nth-child(4)') %>%
    html_text()
  the_max_day_use <- ncLWSP_website %>%
    html_nodes('th~ td+ td') %>%
    html_text()

  #Creating the dataframe
  df_maxdailyuse <- data.frame("Month" = c(1,5,9,2,6,10,3,7,11,4,8,12),
                               "Year" = rep(selected_year,12),
                               "Maximum_Day_Use_MGD" = as.numeric(the_max_day_use)) %>%
  mutate(Water_System_Name = !!the_water_system_name,
         PWSID = !!the_pwsid,
         Ownership = !!the_ownership,
         Date = my(paste(Month,"-",Year)))

  #Adding pause for bulk scraping
  Sys.sleep(1)

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

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

```
#7 Fetching and plotting max daily withdrawals in Durham for each month in 2015

#Fetching max daily withdrawals
df_durham2015 <- scrape.pwsid('03-32-010', 2015)

#Displaying contents of created dataframe
glimpse(df_durham2015)
```

```
## Rows: 12
## Columns: 7
## $ Month               <dbl> 1, 5, 9, 2, 6, 10, 3, 7, 11, 4, 8, 12
## $ Year                <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 20~
## $ Maximum_Day_Use_MGD <dbl> 40.25, 53.17, 40.03, 43.50, 57.02, 38.72, 43.10, 4~
## $ Water_System_Name   <chr> "Durham", "Durham", "Durham", "Durham", "Durham", ~
## $ PWSID               <chr> "03-32-010", "03-32-010", "03-32-010", "03-32-010"~
## $ Ownership           <chr> "Municipality", "Municipality", "Municipality", "M~
## $ Date                <date> 2015-01-01, 2015-05-01, 2015-09-01, 2015-02-01, 20~
```

```
#Creating line plot of the maximum daily withdrawals in Durham for each month in 2015
durham2015.plot <- df_durham2015 %>%
  ggplot(aes(x = Date, y = Maximum_Day_Use_MGD)) +
  geom_line() +
  labs(title = '2015 Maximum Daily Withdrawals per Month in Durham ',
       x = 'Date',
       y = 'Max Day Use (MGD)')

durham2015.plot
```



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 Fetching and plotting max daily withdrawals in Asheville for each month in 2015

#Fetching max daily withdrawals
df_asheville2015 <- scrape.pwsid('01-11-010', 2015)

#Displaying contents of created dataframe
glimpse(df_asheville2015)
```

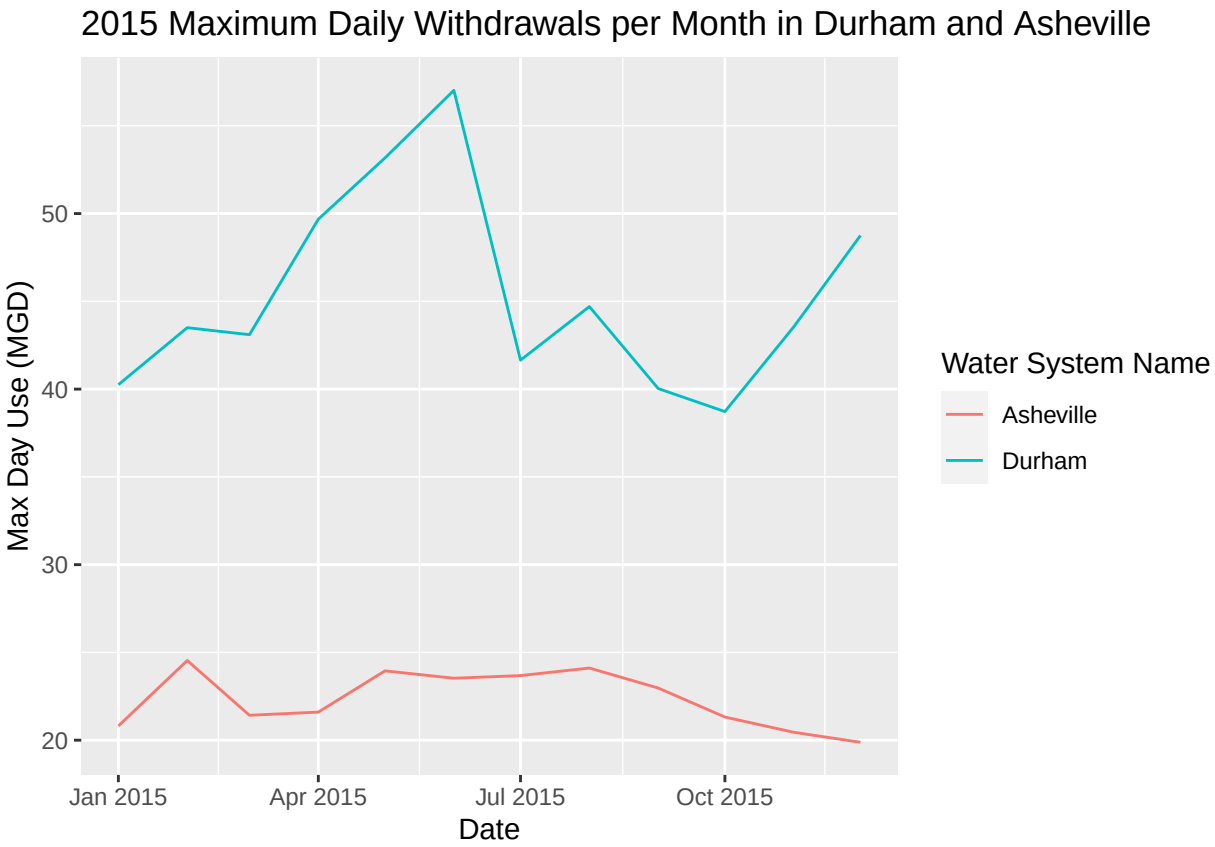
```
## Rows: 12
## Columns: 7
```

```
## $ Month               <dbl> 1, 5, 9, 2, 6, 10, 3, 7, 11, 4, 8, 12
## $ Year                <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 20~
## $ Maximum_Day_Use_MGD <dbl> 20.81, 23.95, 22.97, 24.54, 23.53, 21.32, 21.42, 2~
## $ Water_System_Name   <chr> "Asheville", "Asheville", "Asheville", "Asheville"~
## $ PWSID               <chr> "01-11-010", "01-11-010", "01-11-010", "01-11-010"~
## $ Ownership           <chr> "Municipality", "Municipality", "Municipality", "M~
## $ Date                <date> 2015-01-01, 2015-05-01, 2015-09-01, 2015-02-01, 20~
```

```
#Combining the dataframes
df_durham_asheville2015 <- bind_rows(df_durham2015,df_asheville2015)

#Displaying contents of combined dataframe
glimpse(df_durham_asheville2015)
```

```
## Rows: 24
## Columns: 7
## $ Month               <dbl> 1, 5, 9, 2, 6, 10, 3, 7, 11, 4, 8, 12, 1, 5, 9, 2,~
## $ Year                <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 20~
## $ Maximum_Day_Use_MGD <dbl> 40.25, 53.17, 40.03, 43.50, 57.02, 38.72, 43.10, 4~
## $ Water_System_Name   <chr> "Durham", "Durham", "Durham", "Durham", "Durham", ~
## $ PWSID               <chr> "03-32-010", "03-32-010", "03-32-010", "03-32-010"~
## $ Ownership           <chr> "Municipality", "Municipality", "Municipality", "M~
## $ Date                <date> 2015-01-01, 2015-05-01, 2015-09-01, 2015-02-01, 2~
```

```
durham_asheville2015.plot <- df_durham_asheville2015 %>%
  ggplot(aes(x = Date, y = Maximum_Day_Use_MGD, color = Water_System_Name)) +
  geom_line() +
  labs(title = '2015 Maximum Daily Withdrawals per Month in Durham and Asheville',
       color = 'Water System Name',
       x = 'Date',
       y = 'Max Day Use (MGD)')

durham_asheville2015.plot
```



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.

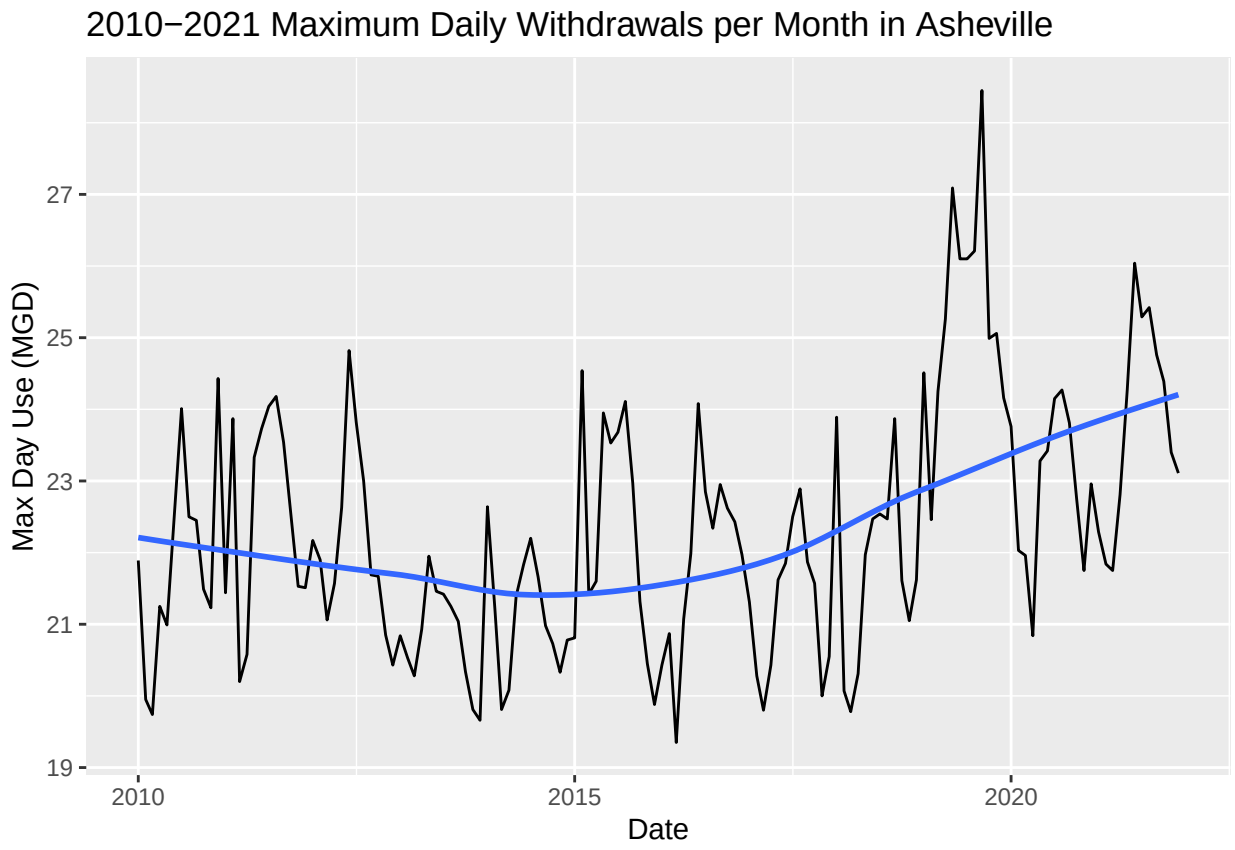
```
#9 Fetching and creating a plot of Asheville's max daily withdrawal by months (2010-2021)

#Selecting Asheville's pwsid
selected_pwsid <- '01-11-010'

#Creating a list of the years desired
sample_years <- rep(2010:2021)

#Using the map2 and scrape_pwsid functions to retrieve data for the selected years
dfs_2010s <- map2(selected_pwsid, sample_years, scrape.pwsid) %>% bind_rows()

#Displaying contents of created dataframe
glimpse(dfs_2010s)
```

```
## Rows: 144
## Columns: 7
## $ Month               <dbl> 1, 5, 9, 2, 6, 10, 3, 7, 11, 4, 8, 12, 1, 5, 9, 2,~
```

```
## $ Year                <int> 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 20~
## $ Maximum_Day_Use_MGD <dbl> 21.89, 20.99, 22.45, 19.95, 22.53, 21.49, 19.74, 2~
## $ Water_System_Name   <chr> "Asheville", "Asheville", "Asheville", "Asheville"~
## $ PWSID               <chr> "01-11-010", "01-11-010", "01-11-010", "01-11-010"~
## $ Ownership           <chr> "Municipality", "Municipality", "Municipality", "M~
## $ Date                <date> 2010-01-01, 2010-05-01, 2010-09-01, 2010-02-01, 2~
```

```
#Creating plot
asheville2010s.plot <- dfs_2010s %>%
  ggplot(aes(x = Date, y = Maximum_Day_Use_MGD)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste('2010-2021 Maximum Daily Withdrawals per Month in Asheville'),
       x = 'Date',
       y = 'Max Day Use (MGD)')

asheville2010s.plot
```

```
## `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: Yes. Based on the plot of maximum daily withdrawal in Asheville by month for the years 2010 - 2021, water usage declined slightly in the years preceding 2015 but has been on an upward trend in the subsequent years.