Assignment 10: Data Scraping

Rosie Wu

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(lubridate)
library(here)
library(rvest)
here()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham's 2023 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=2023

Indicate this website as the as the URL to be scraped. (In other words, read the contents into an `rvest` webpage object.)

```
#2
#Set the URL to be scrapped
# Set the URL
theURL <- 'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2023'

# Read the webpage content
webpage <- read_html(theURL)
```

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
# Scrape/ extract values
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"
```

```
MGD <- webpage %>%
  html_nodes('th~ td+ td , th~ td+ td') %>%
  html_text()
MGD
```

```
##  [1] "28.9000" "33.3000" "43.7000" "30.0000" "40.0000" "37.2300" "34.2000"
##  [8] "44.9000" "40.3500" "30.9000" "56.7000" "33.3000"
```

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 2023, making sure, the months are presented in proper sequence.

```
#4
months <- webpage %>%
  html_nodes('.fancy-table:nth-child(31) tr+ tr th') %>%
  html_text()
months
```

```
##  [1] "Jan" "May" "Sep" "Feb" "Jun" "Oct" "Mar" "Jul" "Nov" "Apr" "Aug" "Dec"
```

```
# Convert max_day_use to numeric
max_day_use <- as.numeric(MGD)
year <- 2023
# Create a Date column, assigning day 1 as the days
dates <- as.Date(paste(year, months, "01", sep = "-"), format = "%Y-%B-%d")

# Create the dataframe
water_data <- data.frame(
  Date = dates,
  Year = rep(year, length(max_day_use)),
  Month = months,
  Water_System_Name = rep(water_system_name, length(max_day_use)),
  PWSID = rep(PWSID, length(max_day_use)),
  Ownership = rep(Ownership, length(max_day_use)),
  Max_Day_Use_MGD = max_day_use
)
print(water_data)
```

```
##          Date Year Month Water_System_Name     PWSID    Ownership
## 1  2023-01-01 2023   Jan            Durham 03-32-010 Municipality
```

```
## 2  2023-05-01 2023   May            Durham 03-32-010 Municipality
## 3  2023-09-01 2023   Sep            Durham 03-32-010 Municipality
## 4  2023-02-01 2023   Feb            Durham 03-32-010 Municipality
## 5  2023-06-01 2023   Jun            Durham 03-32-010 Municipality
## 6  2023-10-01 2023   Oct            Durham 03-32-010 Municipality
## 7  2023-03-01 2023   Mar            Durham 03-32-010 Municipality
## 8  2023-07-01 2023   Jul            Durham 03-32-010 Municipality
## 9  2023-11-01 2023   Nov            Durham 03-32-010 Municipality
## 10 2023-04-01 2023   Apr            Durham 03-32-010 Municipality
## 11 2023-08-01 2023   Aug            Durham 03-32-010 Municipality
## 12 2023-12-01 2023   Dec            Durham 03-32-010 Municipality
##    Max_Day_Use_MGD
## 1            28.90
## 2            33.30
## 3            43.70
## 4            30.00
## 5            40.00
## 6            37.23
## 7            34.20
## 8            44.90
## 9            40.35
## 10           30.90
## 11           56.70
## 12           33.30
```

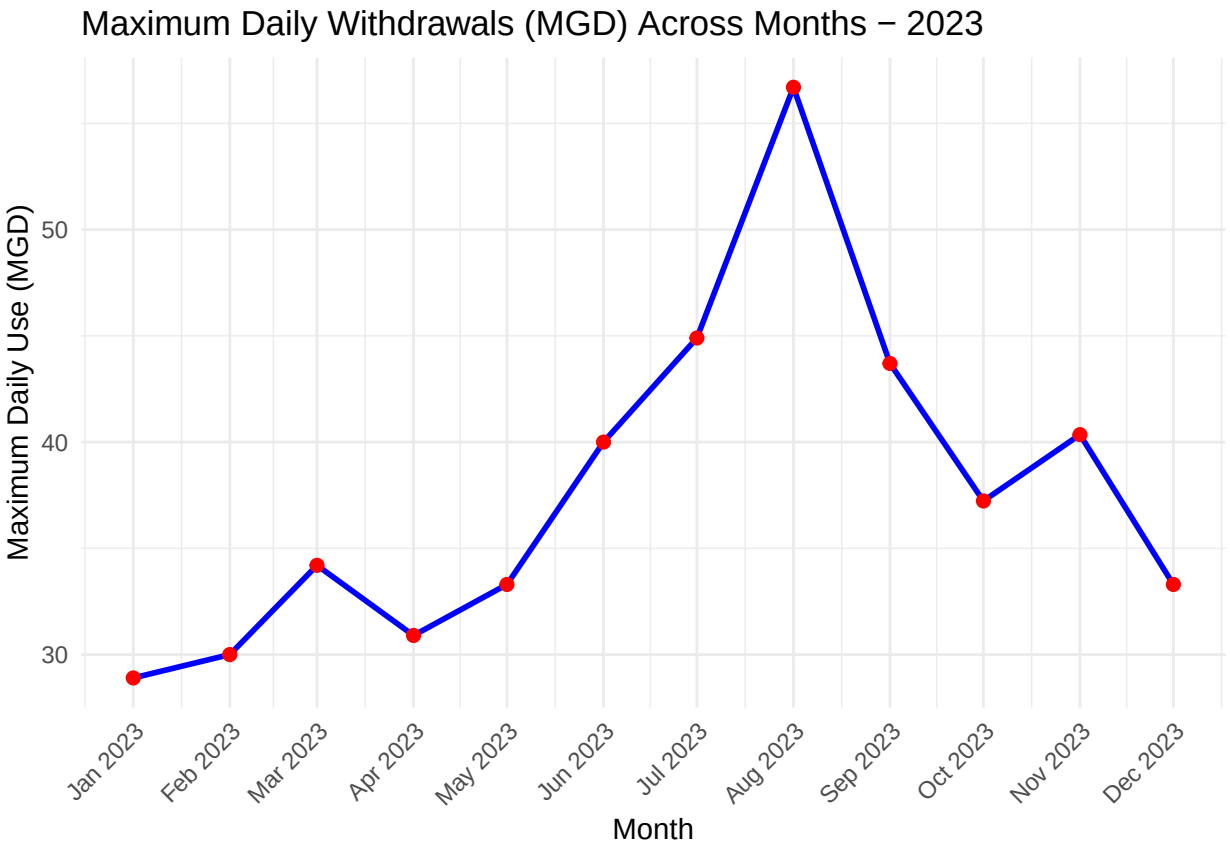
```r
#5
# Sort the rows by the Month column
water_data_sorted <- water_data %>%
  arrange(Date)
print(water_data_sorted)
```

```
##          Date Year Month Water_System_Name     PWSID    Ownership
## 1  2023-01-01 2023   Jan            Durham 03-32-010 Municipality
## 2  2023-02-01 2023   Feb            Durham 03-32-010 Municipality
## 3  2023-03-01 2023   Mar            Durham 03-32-010 Municipality
## 4  2023-04-01 2023   Apr            Durham 03-32-010 Municipality
## 5  2023-05-01 2023   May            Durham 03-32-010 Municipality
## 6  2023-06-01 2023   Jun            Durham 03-32-010 Municipality
## 7  2023-07-01 2023   Jul            Durham 03-32-010 Municipality
## 8  2023-08-01 2023   Aug            Durham 03-32-010 Municipality
## 9  2023-09-01 2023   Sep            Durham 03-32-010 Municipality
## 10 2023-10-01 2023   Oct            Durham 03-32-010 Municipality
## 11 2023-11-01 2023   Nov            Durham 03-32-010 Municipality
## 12 2023-12-01 2023   Dec            Durham 03-32-010 Municipality
##    Max_Day_Use_MGD
## 1            28.90
## 2            30.00
## 3            34.20
## 4            30.90
## 5            33.30
## 6            40.00
## 7            44.90
## 8            56.70
## 9            43.70
```

```
## 10           37.23
## 11           40.35
## 12           33.30
```

```r
library(dplyr)
# Create the line plot
ggplot(water_data_sorted, aes(x = Date, y = Max_Day_Use_MGD, group = 1)) +
  geom_line(color = "blue", size = 1) +
  geom_point(color = "red", size = 2) +
  labs(
    title = "Maximum Daily Withdrawals (MGD) Across Months - 2023",
    x = "Month",
    y = "Maximum Daily Use (MGD)"
  ) +
  scale_x_date(
    date_breaks = "1 month",
    # Force one tick per month, since it wasn't able to include all at first
    date_labels = "%b %Y"             # force abbreviated month and year
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



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, returning a dataframe. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

```
#6. use scrape it function first
scrape.it <- function(PWSID, Year){
  #Get the proper url
 webpage <- read_html(
   paste0(
     'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=', PWSID,
     '&year=', Year))

 #Scrape variables as extracted/ set earlier
 water_system_name <- webpage %>%
  html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>%
  html_text()

 PWSID <- 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()
```

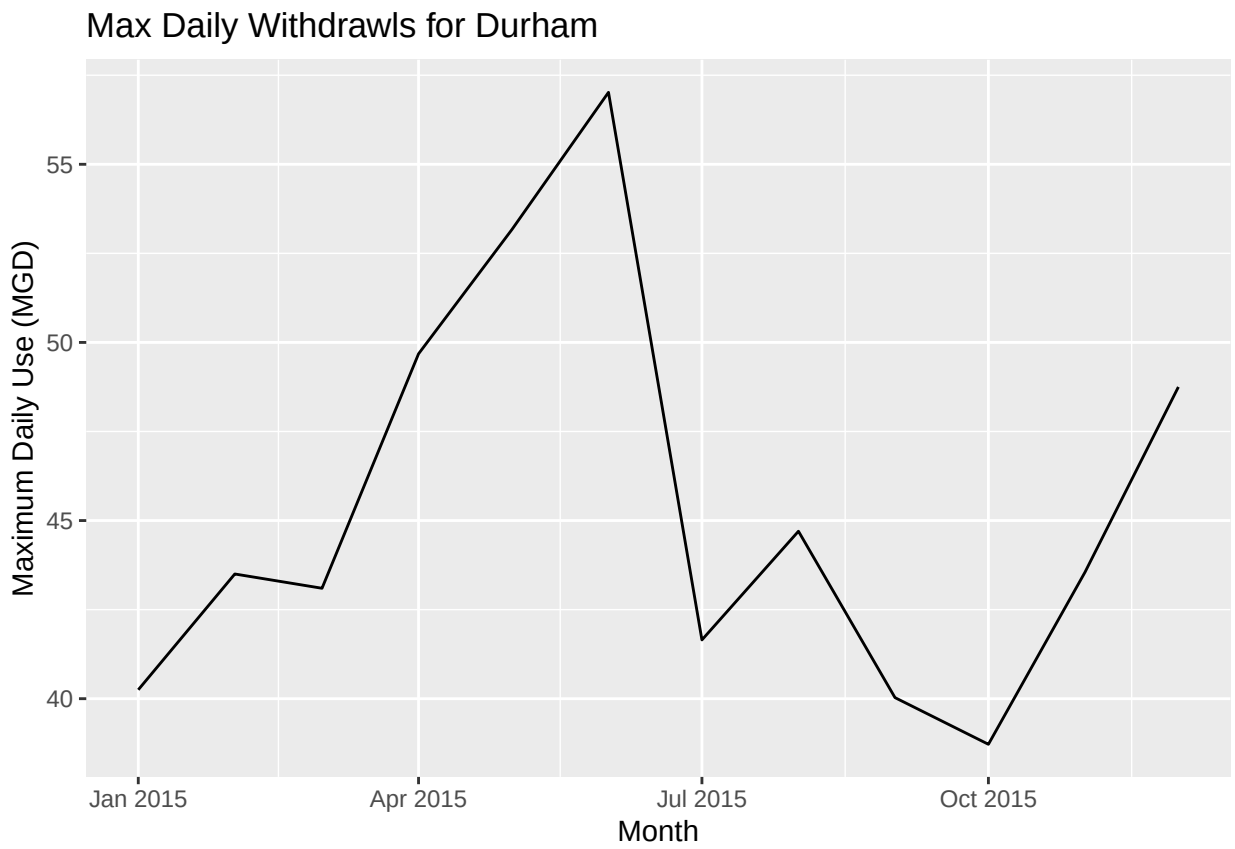
```
MGD <- webpage %>%
 html_nodes('th~ td+ td') %>%
 html_text()

 # Convert and combine everything to data frame
 water_data_6 <- data.frame(
 "Water_System_Name" = rep(water_system_name, 12),
 "PWSID" = rep(PWSID, 12),
 "Ownership" = rep(Ownership, 12),
 "Year" = rep(Year, 12),
 # Have to redefine this since it's different for Ashville
 "Month" = c("Jan", "May", "Sept", "Feb", "Jun", "Oct", "Mar", "Jul", "Nov",
             "Apr", "Aug", "Dec"),
 "MGD" = as.numeric(gsub(',','', MGD )))%>%
   mutate("Date" = paste(Month, Year, sep = "-"))
 return(water_data_6)
 }
```

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

```
#7
water_withdraw_Durham_2015 <- scrape.it('03-32-010', 2015)
water_withdraw_Durham_2015
```

```
##    Water_System_Name     PWSID    Ownership Year Month   MGD      Date
## 1             Durham 03-32-010 Municipality 2015   Jan 40.25  Jan-2015
## 2             Durham 03-32-010 Municipality 2015   May 53.17  May-2015
## 3             Durham 03-32-010 Municipality 2015  Sept 40.03 Sept-2015
## 4             Durham 03-32-010 Municipality 2015   Feb 43.50  Feb-2015
## 5             Durham 03-32-010 Municipality 2015   Jun 57.02  Jun-2015
## 6             Durham 03-32-010 Municipality 2015   Oct 38.72  Oct-2015
## 7             Durham 03-32-010 Municipality 2015   Mar 43.10  Mar-2015
## 8             Durham 03-32-010 Municipality 2015   Jul 41.65  Jul-2015
## 9             Durham 03-32-010 Municipality 2015   Nov 43.55  Nov-2015
## 10            Durham 03-32-010 Municipality 2015   Apr 49.68  Apr-2015
## 11            Durham 03-32-010 Municipality 2015   Aug 44.70  Aug-2015
## 12            Durham 03-32-010 Municipality 2015   Dec 48.75  Dec-2015
```

```
# Example plot
# first convert to date format
water_withdraw_Durham_2015$Date <- my(water_withdraw_Durham_2015$Date)
# plot it
water_withdraw_Durham_2015 %>%
  ggplot(aes(x=Date, y=MGD)) +
  geom_line()+
  labs(title = "Max Daily Withdrawls for Durham",
       x = "Month",
    y = "Maximum Daily Use (MGD)")
```

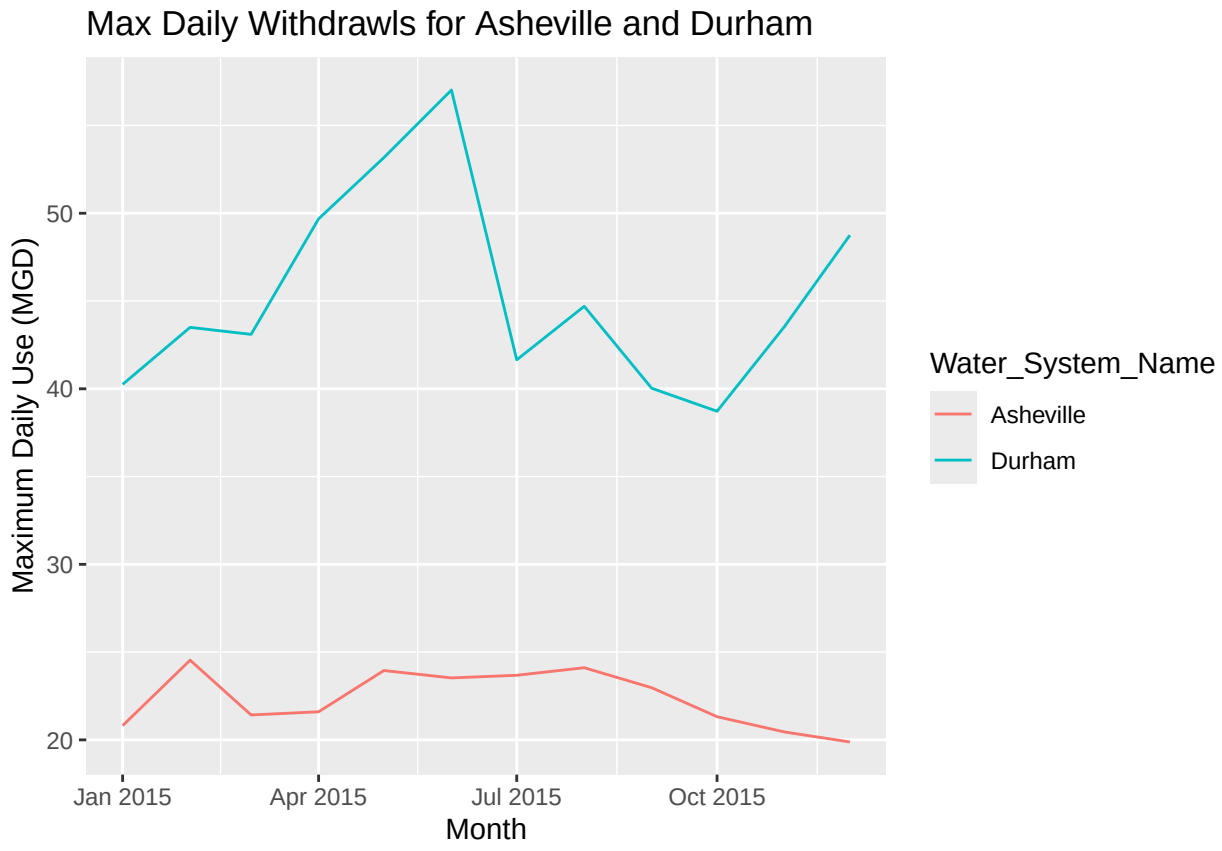


8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares Asheville's to Durham's water withdrawals.

```
#8
# scrape data and define function first
water_withdraw_Asheville_2015 <- scrape.it('01-11-010', 2015)
water_withdraw_Asheville_2015
```

```
##    Water_System_Name     PWSID    Ownership Year Month   MGD      Date
## 1          Asheville 01-11-010 Municipality 2015   Jan 20.81  Jan-2015
## 2          Asheville 01-11-010 Municipality 2015   May 23.95  May-2015
## 3          Asheville 01-11-010 Municipality 2015  Sept 22.97 Sept-2015
## 4          Asheville 01-11-010 Municipality 2015   Feb 24.54  Feb-2015
## 5          Asheville 01-11-010 Municipality 2015   Jun 23.53  Jun-2015
## 6          Asheville 01-11-010 Municipality 2015   Oct 21.32  Oct-2015
## 7          Asheville 01-11-010 Municipality 2015   Mar 21.42  Mar-2015
## 8          Asheville 01-11-010 Municipality 2015   Jul 23.68  Jul-2015
## 9          Asheville 01-11-010 Municipality 2015   Nov 20.45  Nov-2015
## 10         Asheville 01-11-010 Municipality 2015   Apr 21.60  Apr-2015
## 11         Asheville 01-11-010 Municipality 2015   Aug 24.11  Aug-2015
## 12         Asheville 01-11-010 Municipality 2015   Dec 19.88  Dec-2015
```

```r
#Convert to date format
water_withdraw_Asheville_2015$Date <- my(water_withdraw_Asheville_2015$Date)

# create a combined dataframe for Asheville and Durham
combined_set <- rbind(water_withdraw_Asheville_2015, water_withdraw_Durham_2015)

# plot the combined data
combined_set %>%
  ggplot(aes(x=Date, y=MGD, color = Water_System_Name)) +
  geom_line()+
  labs(title = "Max Daily Withdrawls for Asheville and Durham",
       x = "Month",
       y = "Maximum Daily Use (MGD)",
       color = "Water_System_Name")
```



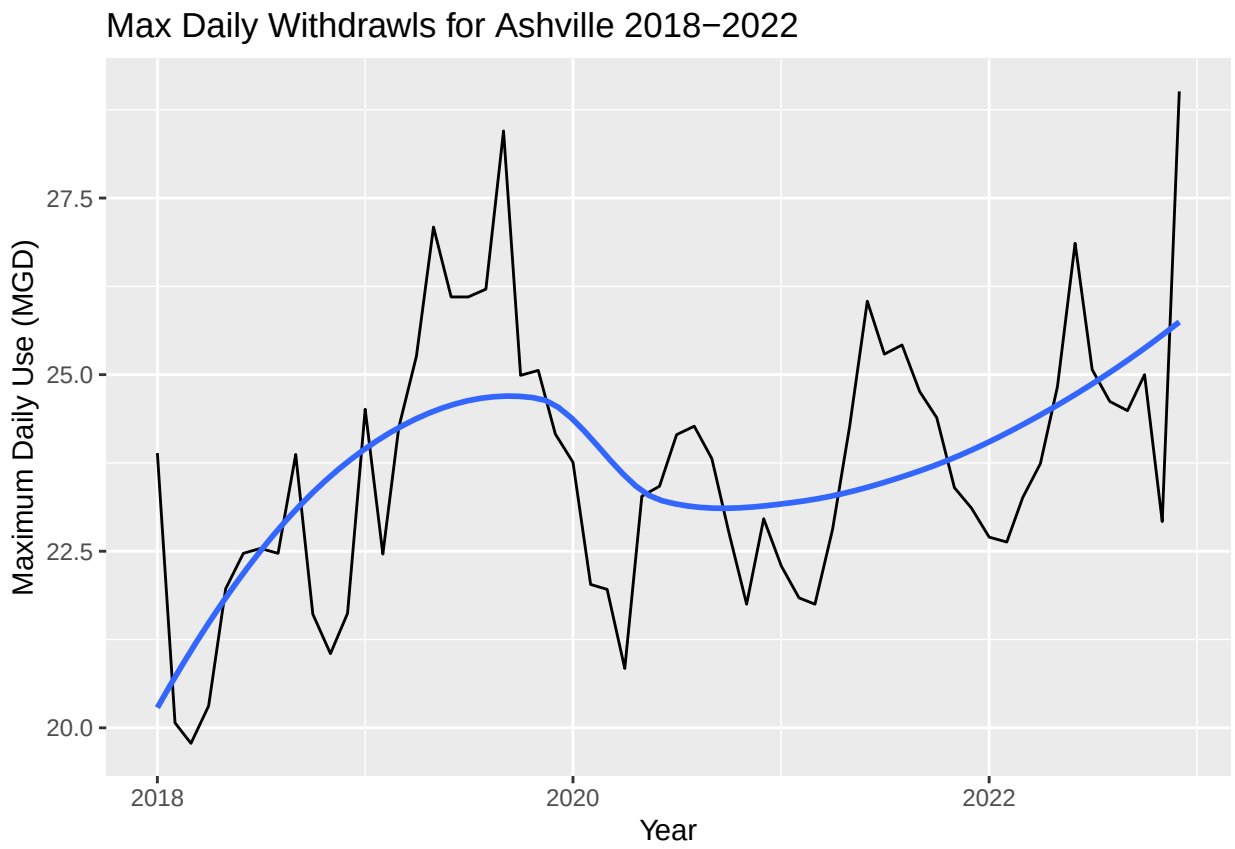
9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2018 thru 2022.Add a smoothed line to the plot (method = 'loess').

   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.

```r
#9
the_years <- c(2018:2022)
```

```
Ash_years_combined <- map2('01-11-010',the_years,scrape.it)
# Asheville over years
Ash_years_combined <- bind_rows(Ash_years_combined)
Ash_years_combined$Date <- my(Ash_years_combined$Date)
#Plot
Ash_years_combined %>%
  ggplot(aes(x=Date, y=MGD)) +
  geom_line()+
  geom_smooth(method="loess",se=FALSE)+
  labs(title = "Max Daily Withdrawls for Ashville 2018-2022",
       x = "Year",
    y = "Maximum Daily Use (MGD)")
```

```
## `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: > In general, the water use over time has a increasing trend over year 2018-2022. There is a decrease/ drop right after year 2020, this can potentially be due to decrease in general activities around the time of Covid. Meanwhile, if we were to focus on trend within each year, the water use is highest around late summer or early fall, which is when resdients tend to have more outdoor activities and shower more to use more water.