Assignment 10: Data Scraping

Aishwarya Patankar

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); here()
```

```
## [1] "/Users/aishwaryapatankar/Documents/Duke University/Spring2025/ENERGY872/EDA_Spring2025"
```

```
#install.packages("rvest")
library(rvest)
```

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham's 2024 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=2024

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

```
#2
theURL <- read_html("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2024")
```

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
max_withdrawals <- theURL %>%
  html_nodes("th~ td+ td")%>%
  html_text()
max_withdrawals
```

```
##  [1] "34.5000" "36.0600" "37.3300" "32.1000" "46.6500" "37.3600" "38.2000"
##  [8] "41.9000" "36.5800" "36.7300" "42.9600" "34.4500"
```

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

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

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

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

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

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

```r
the_Month <- theURL %>%
  html_nodes(".fancy-table:nth-child(30) tr+ tr th")%>%
  html_text()
the_Month
```

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

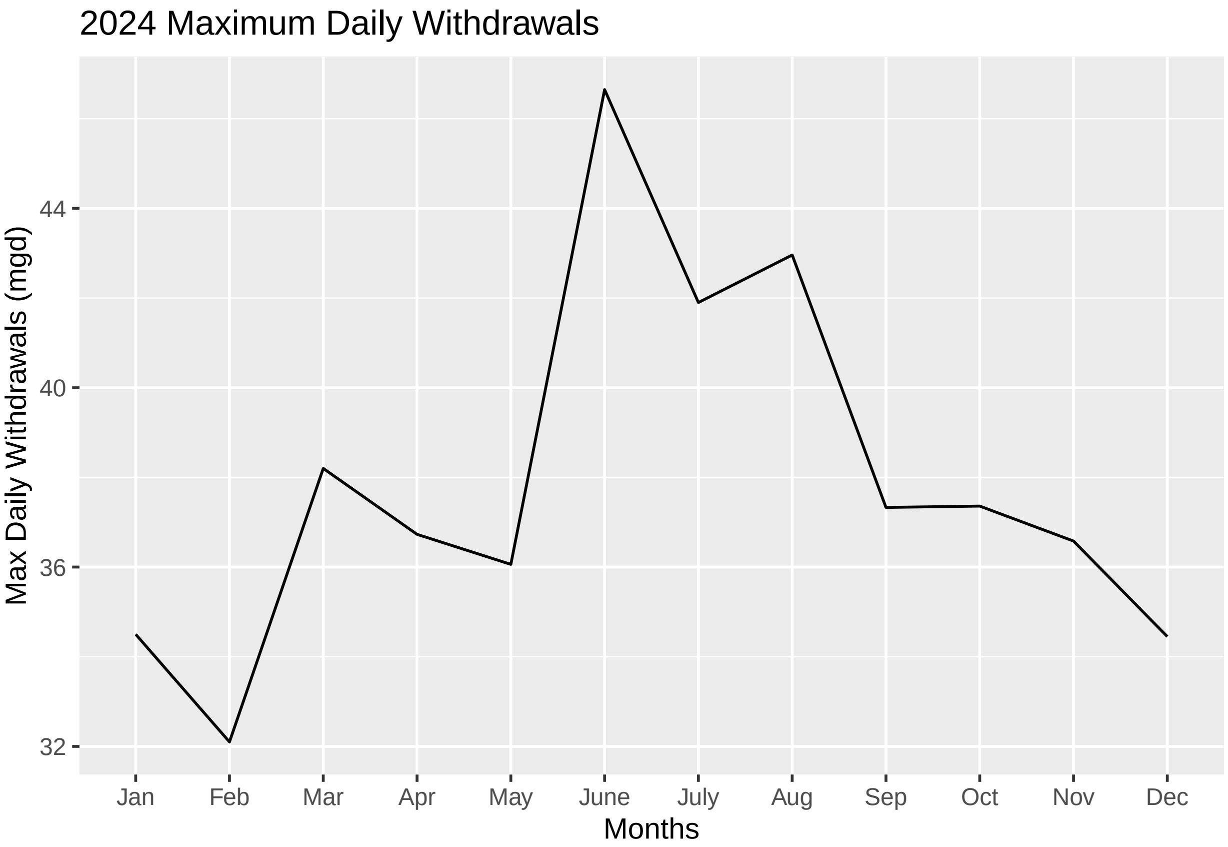
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   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 2024, making sure, the months are presented in proper sequence.

```r
#4
df_withdrawals <- data.frame( "Year" = rep(2024,12),
                             "Max_withdrawals_mgd" = as.numeric(max_withdrawals))

month_order <- c("Jan","May","Sep","Feb","Jun","Oct","Mar","Jul","Nov","Apr","Aug","Dec")
month_to_number <- c("Jan"="01","May"="05","Sep"="09","Feb"="02","Jun"="06","Oct"="10","Mar"="03","Jul"=

#Modifying the Dataframe to include Water System Owner name, PWSID and Ownership
df_withdrawals <- df_withdrawals %>%
  mutate(WaterSystemName = !!the_WaterSystemName,
         PWSID = !!the_PWSID,
         Municipality = !!the_municipality,
         Month = month_order,
         Date = as.Date(paste(df_withdrawals$Year,month_to_number[Month], "01", sep ="-")))

df_withdrawals_arranged <- df_withdrawals %>%
  arrange(Date)%>%
  mutate(Month_new = month(Date))%>%
  mutate(Month_new = factor(Month_new, levels = 1:12, labels = c("Jan", "Feb", "Mar", "Apr", "May", "Ju

#5
Plot_MaxDailyWithdrawals <-
  ggplot(df_withdrawals_arranged, aes(x=Month_new, y=Max_withdrawals_mgd))+
  geom_line(group = 1)+
  labs(title = "2024 Maximum Daily Withdrawals", x = "Months", y ="Max Daily Withdrawals (mgd)")
  print(Plot_MaxDailyWithdrawals)
```



6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function with two input - "PWSID" and "year" - that:

- Creates a URL pointing to the LWSP for that PWSID for the given year
- Creates a website object and scrapes the data from that object (just as you did above)
- Constructs a dataframe from the scraped data, mostly as you did above, but includes the PWSID and year provided as function inputs in the dataframe.
- Returns the dataframe as the function's output

```
#6.
#Constructing the scraping URL
the_BaseURL <- 'https://www.ncwater.org/WUDC/app/LWSP/report.php?'
Q6the_PWSID <- '03-32-010'
the_year <- 2020
the_scrape_url <- paste0(the_BaseURL,'pwsid=',Q6the_PWSID,'&year=',the_year)
print(the_scrape_url)
```

```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2020"
```

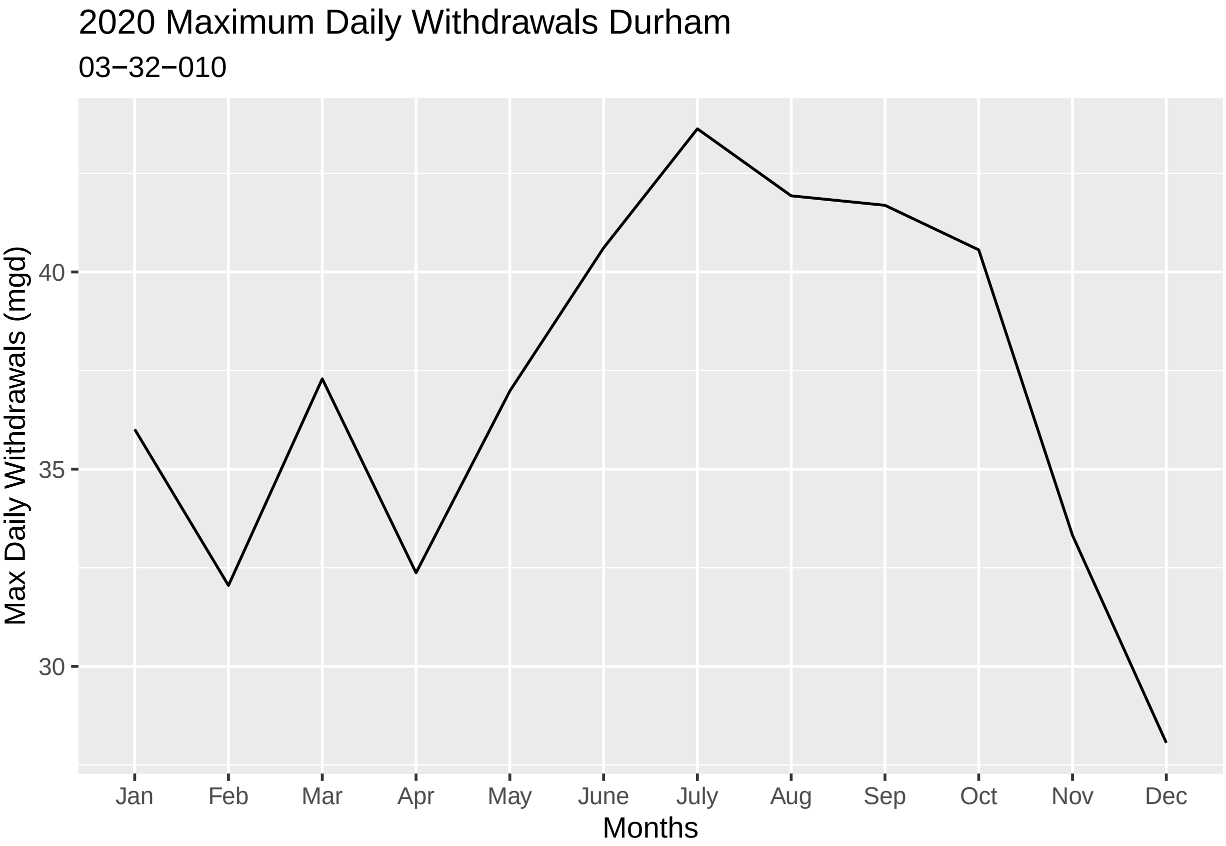
```
# Retrieving website contents
the_website <- read_html(the_scrape_url)

#Set the element address variables

max_withdrawals_tag <- 'th~ td+ td'
```

```
the_municipality_tag <-'div+ table tr:nth-child(2) td:nth-child(4)'
the_PWSID_tag <- 'td tr:nth-child(1) td:nth-child(5)'
the_WaterSystemName_tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'

#Scraping the data items
max_withdrawals_data <- the_website %>% html_nodes(max_withdrawals_tag) %>%html_text()
the_municipality_data <- the_website %>% html_nodes(the_municipality_tag) %>%html_text()
the_PWSID_data <- the_website %>% html_nodes(the_PWSID_tag) %>%html_text()
the_WaterSystemName_data <- the_website %>% html_nodes(the_WaterSystemName_tag) %>%html_text()

#Converting to a dataframe
df_withdrawals2 <- data.frame( "Year" = rep(the_year,12),
                             "Max_withdrawals_mgd" = as.numeric(max_withdrawals_data))

#Modifying the Dataframe to include Water System Owner name, PWSID and Ownership
df_withdrawals2 <- df_withdrawals2 %>%
  mutate(WaterSystemName = !!the_WaterSystemName_data,
         PWSID = !!the_PWSID_data,
         Municipality = !!the_municipality_data,
         Month = month_order,
         Date = as.Date(paste(df_withdrawals$Year,month_to_number[Month], "01", sep ="-")))

df_withdrawals_arranged2 <- df_withdrawals2 %>%
  arrange(Date)%>%
  mutate(Month_new = month(Date))%>%
  mutate(Month_new = factor(Month_new, levels = 1:12, labels = c("Jan", "Feb", "Mar", "Apr", "May", "Ju
return(df_withdrawals_arranged2)
```

```
##    Year Max_withdrawals_mgd WaterSystemName     PWSID Municipality Month
## 1  2020               36.01          Durham 03-32-010 Municipality   Jan
## 2  2020               32.05          Durham 03-32-010 Municipality   Feb
## 3  2020               37.29          Durham 03-32-010 Municipality   Mar
## 4  2020               32.37          Durham 03-32-010 Municipality   Apr
## 5  2020               36.98          Durham 03-32-010 Municipality   May
## 6  2020               40.61          Durham 03-32-010 Municipality   Jun
## 7  2020               43.63          Durham 03-32-010 Municipality   Jul
## 8  2020               41.93          Durham 03-32-010 Municipality   Aug
## 9  2020               41.69          Durham 03-32-010 Municipality   Sep
## 10 2020               40.56          Durham 03-32-010 Municipality   Oct
## 11 2020               33.32          Durham 03-32-010 Municipality   Nov
## 12 2020               28.06          Durham 03-32-010 Municipality   Dec
##          Date Month_new
## 1  2024-01-01       Jan
## 2  2024-02-01       Feb
## 3  2024-03-01       Mar
## 4  2024-04-01       Apr
## 5  2024-05-01       May
## 6  2024-06-01      June
## 7  2024-07-01      July
## 8  2024-08-01       Aug
## 9  2024-09-01       Sep
## 10 2024-10-01       Oct
## 11 2024-11-01       Nov
## 12 2024-12-01       Dec
```

7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID='03-32-010') for each month in 2020

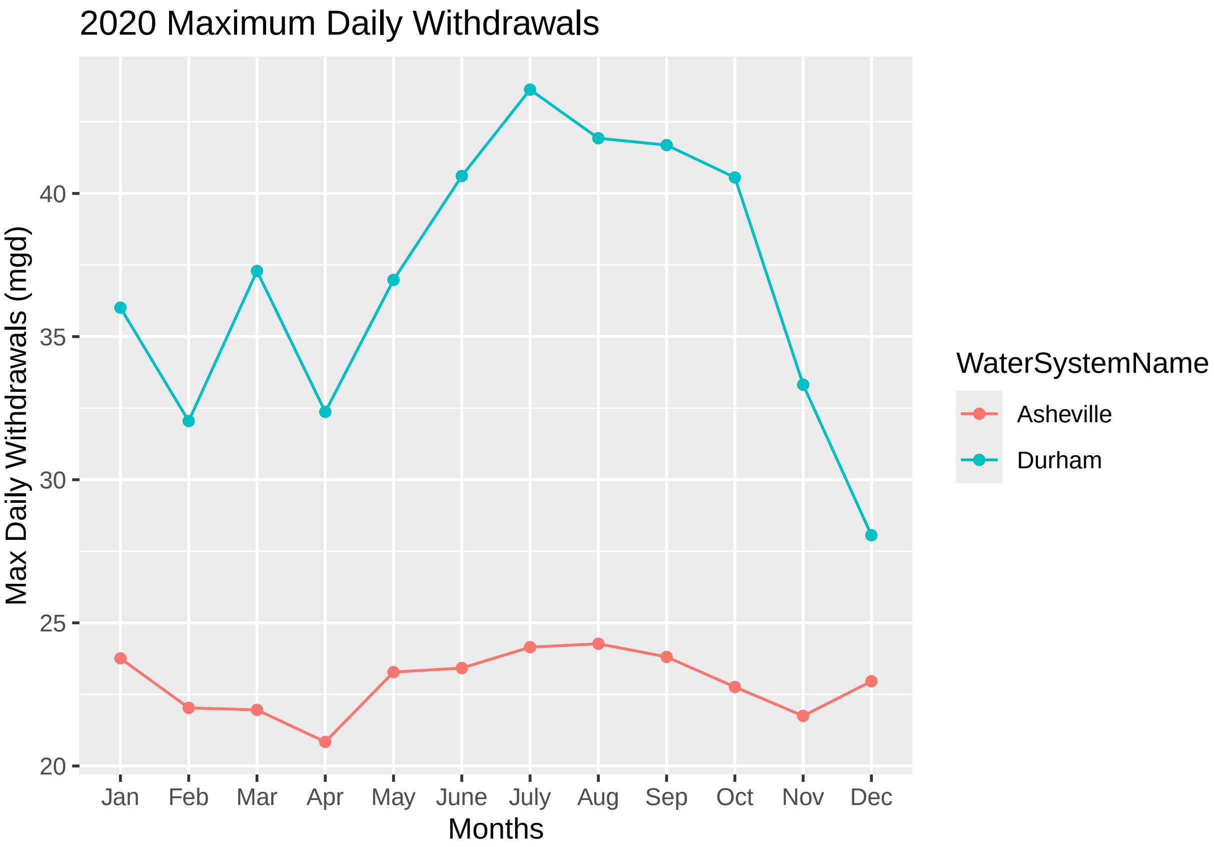
```
#7
Plot_MaxDailyWithdrawals2 <-
  ggplot(df_withdrawals_arranged2, aes(x=Month_new, y=Max_withdrawals_mgd))+
  geom_line(group = 1)+
  labs(title = paste(the_year, "Maximum Daily Withdrawals", the_WaterSystemName_data), subtitle = paste
  print(Plot_MaxDailyWithdrawals2)
```



8. Use the function above to extract data for Asheville (PWSID = '01-11-010') in 2020. Combine this data with the Durham data collected above and create a plot that compares Asheville's to Durham's water withdrawals.

```
#8

#Creating a Scraping function
scrape.it <- function(the_year,Q6the_PWSID){
  #Constructing the scraping URL
the_BaseURL <-  'https://www.ncwater.org/WUDC/app/LWSP/report.php?'
the_scrape_url <- paste0(the_BaseURL,'pwsid=',Q6the_PWSID,'&year=',the_year)

# Retrieving website contents
the_website <- read_html(the_scrape_url)
```

```
#Set the element address variables

max_withdrawals_tag <- 'th~ td+ td'
the_municipality_tag <-'div+ table tr:nth-child(2) td:nth-child(4)'
the_PWSID_tag <- 'td tr:nth-child(1) td:nth-child(5)'
the_WaterSystemName_tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'

#Scraping the data items
max_withdrawals_data <- the_website %>% html_nodes(max_withdrawals_tag) %>%html_text()
the_municipality_data <- the_website %>% html_nodes(the_municipality_tag) %>%html_text()
the_PWSID_data <- the_website %>% html_nodes(the_PWSID_tag) %>%html_text()
the_WaterSystemName_data <- the_website %>% html_nodes(the_WaterSystemName_tag) %>%html_text()

#Converting to a dataframe
df_withdrawals2 <- data.frame( "Year" = rep(the_year,12),
                             "Max_withdrawals_mgd" = as.numeric(max_withdrawals_data))

#Modifying the Dataframe to include Water System Owner name, PWSID and Ownership
df_withdrawals2 <- df_withdrawals2 %>%
  mutate(WaterSystemName = !!the_WaterSystemName_data,
         PWSID = !!the_PWSID_data,
         Municipality = !!the_municipality_data,
         Month = month_order,
         Date = as.Date(paste(df_withdrawals$Year,month_to_number[Month], "01", sep ="-")))

df_withdrawals_arranged2 <- df_withdrawals2 %>%
  arrange(Date)%>%
  mutate(Month_new = month(Date))%>%
  mutate(Month_new = factor(Month_new, levels = 1:12, labels = c("Jan", "Feb", "Mar", "Apr", "May", "Ju

return(df_withdrawals_arranged2)
}

Ashville_2020 <- scrape.it(2020,'01-11-010')

Dfs2020 <- bind_rows(df_withdrawals_arranged2,Ashville_2020)

Dfs2020 <- Dfs2020 %>% arrange(Month_new)

Plot_2020 <-
  ggplot(Dfs2020, aes(x=Month_new, y=Max_withdrawals_mgd, color = WaterSystemName, group = WaterSystemNa
  geom_line()+
  geom_point()+
  labs(title = paste(the_year, "Maximum Daily Withdrawals"), x = "Months", y ="Max Daily Withdrawals (mg
  print(Plot_2020)
```

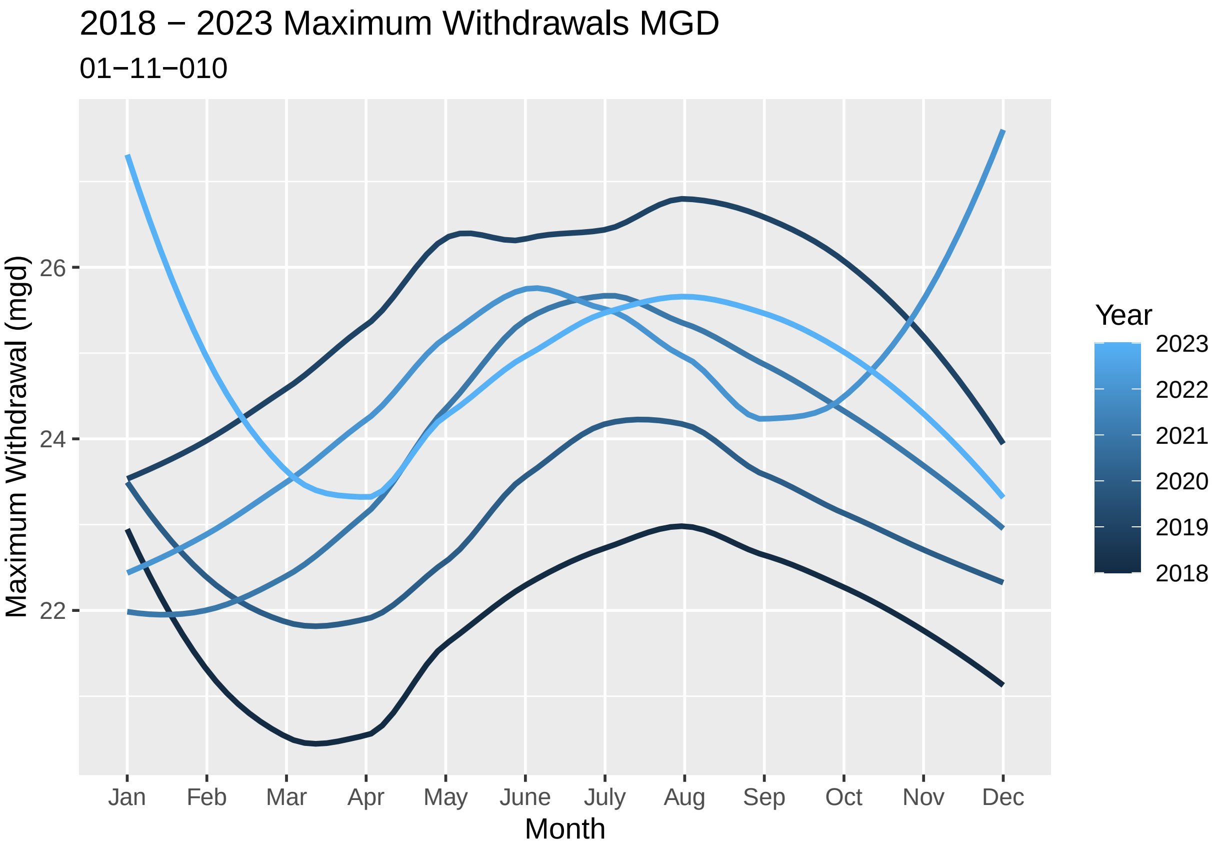


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2018 thru 2023.Add a smoothed line to the plot (method = 'loess').

   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, and use that to construct your plot.

```
#9
the_years = rep(2018:2023)
myfacility = '01-11-010'

the_dfs <- map(the_years, scrape.it, Q6the_PWSID =myfacility)

the_df <- bind_rows(the_dfs)

#Plotting
Ashville_Plot<-
ggplot(the_df, aes(x = Month_new, y=Max_withdrawals_mgd, color=Year, group=Year))+
  geom_smooth(method="loess",se=FALSE)+
  labs(title = paste("2018 - 2023 Maximum Withdrawals MGD"), subtitle = myfacility, x="Month", y="Maximu
print(Ashville_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, by looking at the plot we see that in summer months typically from May to August the water demand is high. The demand reduces in the winter months and some dips in demand are also observed in Spring. Some outliers are also observed i.e a high water demand is observed in the winter of 2022 and beginning of 2023.