Assignment 10: Data Scraping

Vignesh Arunkumar

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)

getwd()
```

```
## [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
lwsp.url <-
"https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2023"

lwsp_durham <- read_html(lwsp.url)
```

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
water_system_name <- lwsp_durham %>%
  html_element("td tr:nth-child(1) td:nth-child(2)") %>%
  html_text(trim = TRUE)

pwsid <- lwsp_durham %>%
  html_element("td tr:nth-child(1) td:nth-child(5)") %>%
  html_text(trim = TRUE)

ownership <-  lwsp_durham %>%
  html_element("tr:nth-child(2) td:nth-child(4)") %>%
  html_text(trim = TRUE)

max_day_use <- lwsp_durham %>%
  html_elements("th~ td+ td") %>%
  html_text(trim = TRUE)
```

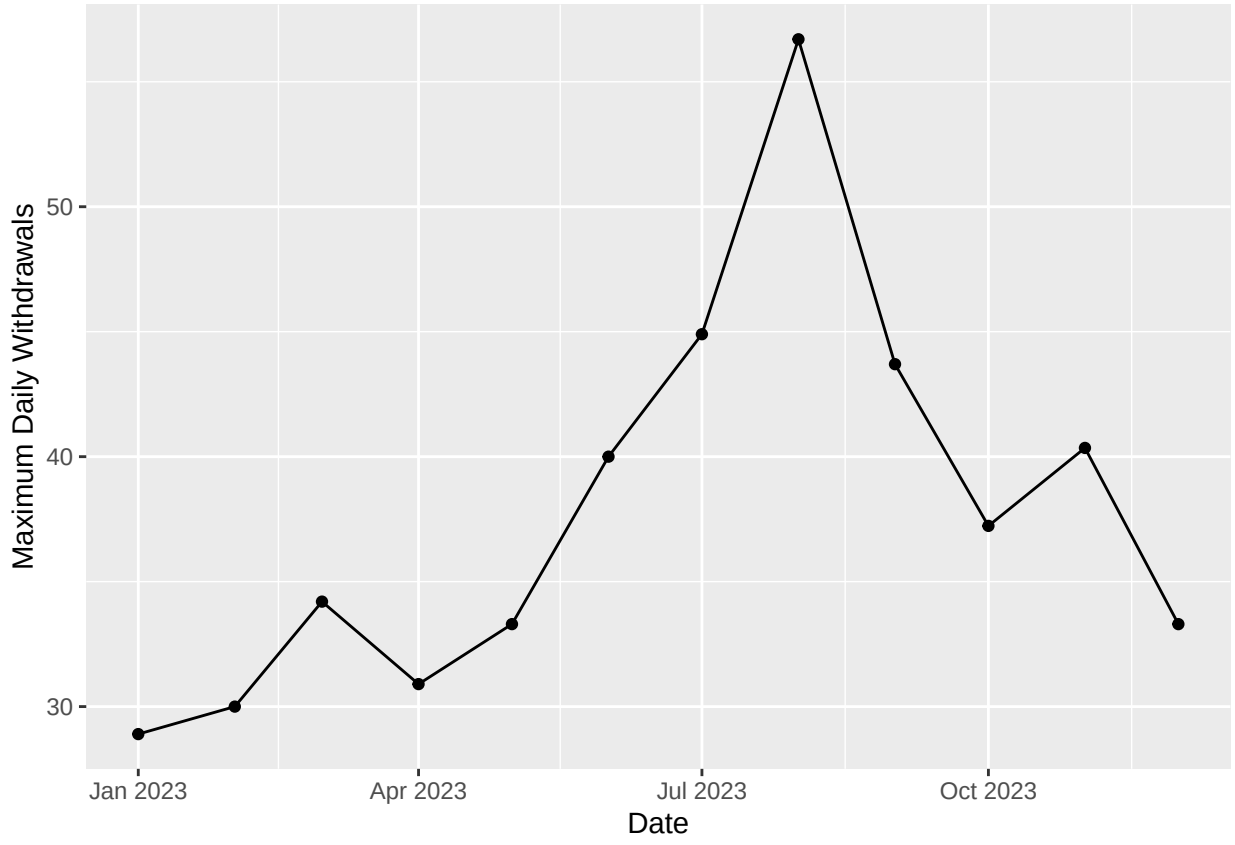
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   > 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 <- c("Jan" ,"May", "Sept", "Feb", "Jun", "Oct", "Mar", "Jul", "Nov",
            "Apr", "Aug", "Dec")
months.num <- c(01, 05, 09, 02, 06, 10, 03, 07, 11, 04, 08, 12)
years <- rep(2023, 12)
lwsp.dataframe <- data.frame("Water System Name" = water_system_name,
                             "PWSID" = pwsid,
                             "Ownership" = ownership,
                             "Max Day Use" = as.numeric(max_day_use) ,
                             "Month" = months,
                             "Date"= make_date(years, months.num))
lwsp.dataframe <- arrange(lwsp.dataframe, Date)

#5
ggplot(lwsp.dataframe, aes(x= Date, y= Max.Day.Use)) +
  geom_point()+
  geom_line()+
labs(x = "Date",y = "Maximum Daily Withdrawals")
```



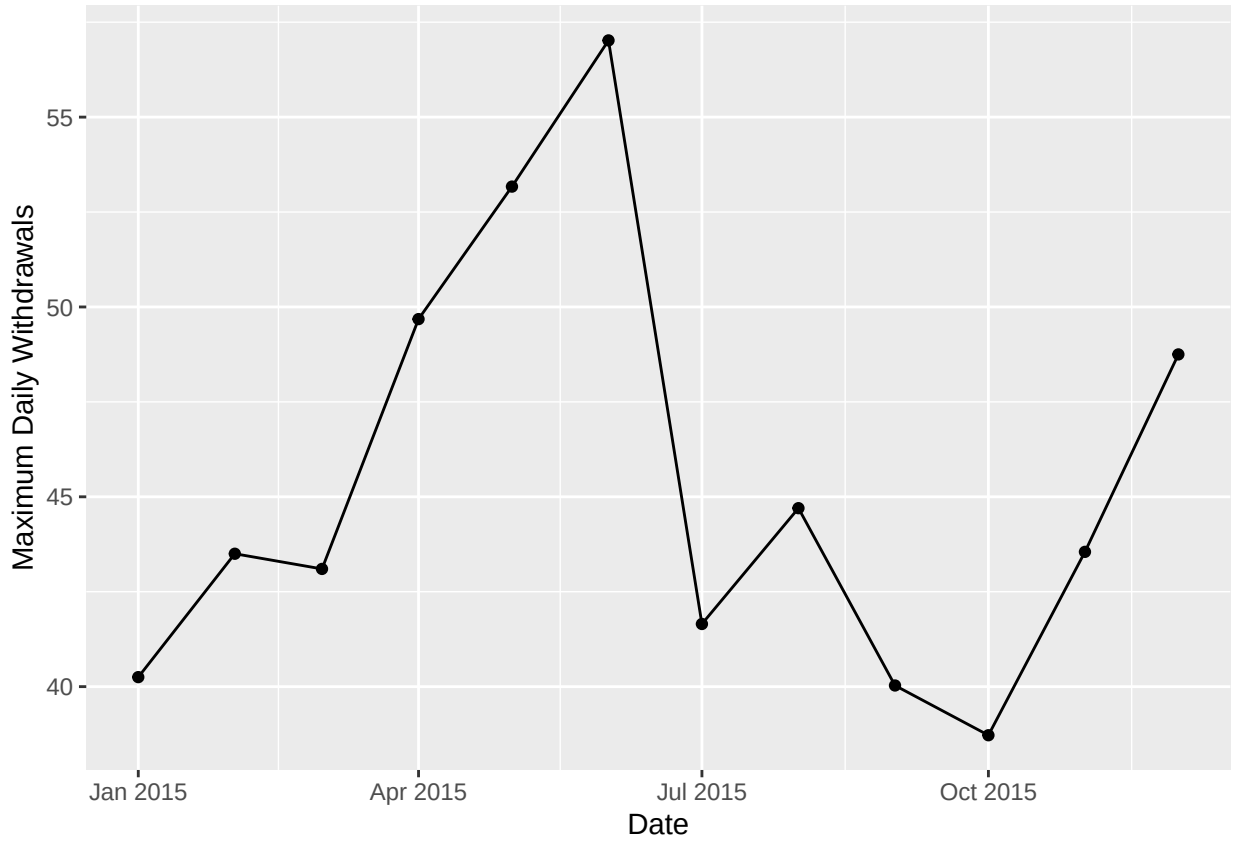
6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data, returning a dataframe. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

```
#6.
scrape.it <- function(ID, Yr){
  url <-
gsub(" ","",paste("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=",
                  ID,"&year=",Yr))
read.my.url <- read_html(url)
print(url)
#scrapedata
system_name <- read.my.url %>%
  html_element("td tr:nth-child(1) td:nth-child(2)") %>%
  html_text(trim = TRUE)

the.id <- read.my.url %>%
  html_element("td tr:nth-child(1) td:nth-child(5)") %>%
  html_text(trim = TRUE)

owner <-  read.my.url %>%
  html_element("tr:nth-child(2) td:nth-child(4)") %>%
  html_text(trim = TRUE)

maxuseday <- read.my.url %>%
  html_elements("th~ td+ td")%>%
  html_text(trim = TRUE)



#creates dataframe
months <- c("Jan" ,"May", "Sept", "Feb", "Jun", "Oct", "Mar", "Jul", "Nov",
            "Apr", "Aug", "Dec")
months.num <- c(01, 05, 09, 02, 06, 10, 03, 07, 11, 04, 08, 12)
years <- rep(Yr, 12)
df <- data.frame("Water System Name" = system_name,
                              "PWSID" = the.id,
                              "Ownership" = owner,
                              "Max Day Use" = as.numeric(maxuseday) ,
                              "Month" = months,
                              "Date"= make_date(years, months.num))


#plot daily use
plot <- ggplot(df, aes(x= Date, y= Max.Day.Use)) +
  geom_point()+
  geom_line()+
labs(x = "Date",y = "Maximum Daily Withdrawals")
print(plot)

#return dataframe
df<- arrange(df, Date)

print(df)
return(df)
}
```

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

```
#7
durham15 <- scrape.it('03-32-010', '2015')
```

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

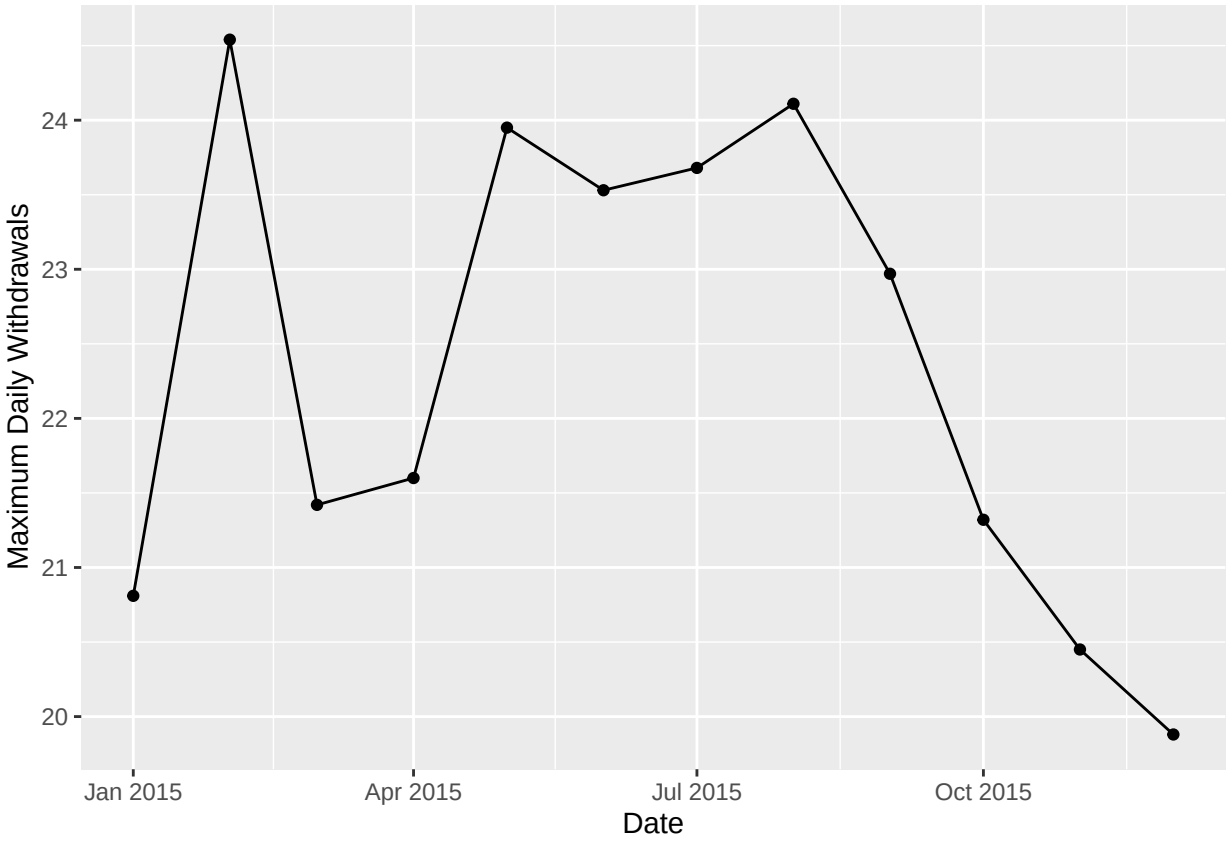



```
##    Water.System.Name     PWSID    Ownership Max.Day.Use Month       Date
## 1            Durham 03-32-010 Municipality       40.25   Jan 2015-01-01
## 2            Durham 03-32-010 Municipality       43.50   Feb 2015-02-01
## 3            Durham 03-32-010 Municipality       43.10   Mar 2015-03-01
## 4            Durham 03-32-010 Municipality       49.68   Apr 2015-04-01
## 5            Durham 03-32-010 Municipality       53.17   May 2015-05-01
## 6            Durham 03-32-010 Municipality       57.02   Jun 2015-06-01
## 7            Durham 03-32-010 Municipality       41.65   Jul 2015-07-01
## 8            Durham 03-32-010 Municipality       44.70   Aug 2015-08-01
## 9            Durham 03-32-010 Municipality       40.03  Sept 2015-09-01
## 10           Durham 03-32-010 Municipality       38.72   Oct 2015-10-01
## 11           Durham 03-32-010 Municipality       43.55   Nov 2015-11-01
## 12           Durham 03-32-010 Municipality       48.75   Dec 2015-12-01
```

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.

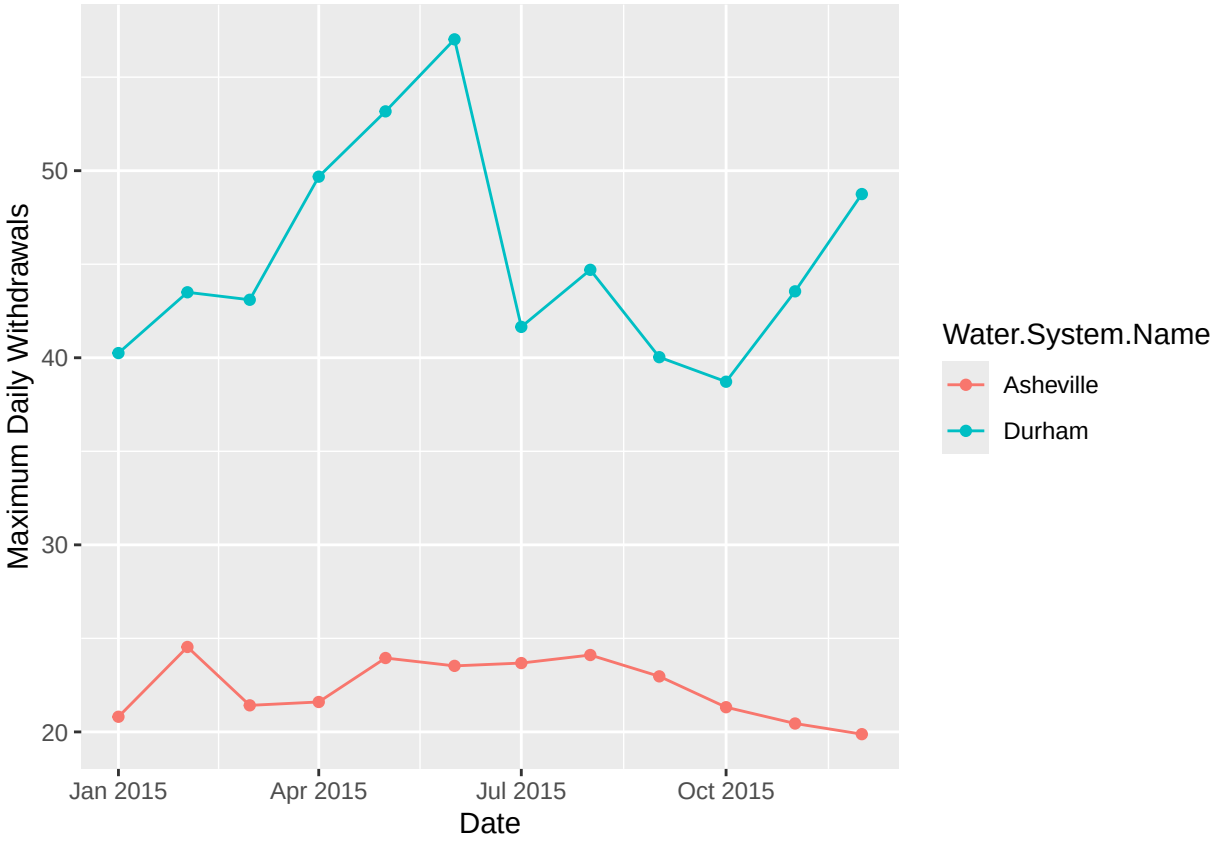
```
#8
ash <- scrape.it('01-11-010', '2015')
```

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




```
##    Water.System.Name     PWSID    Ownership Max.Day.Use Month       Date
## 1          Asheville 01-11-010 Municipality       20.81   Jan 2015-01-01
## 2          Asheville 01-11-010 Municipality       24.54   Feb 2015-02-01
## 3          Asheville 01-11-010 Municipality       21.42   Mar 2015-03-01
## 4          Asheville 01-11-010 Municipality       21.60   Apr 2015-04-01
## 5          Asheville 01-11-010 Municipality       23.95   May 2015-05-01
## 6          Asheville 01-11-010 Municipality       23.53   Jun 2015-06-01
## 7          Asheville 01-11-010 Municipality       23.68   Jul 2015-07-01
## 8          Asheville 01-11-010 Municipality       24.11   Aug 2015-08-01
## 9          Asheville 01-11-010 Municipality       22.97  Sept 2015-09-01
## 10         Asheville 01-11-010 Municipality       21.32   Oct 2015-10-01
## 11         Asheville 01-11-010 Municipality       20.45   Nov 2015-11-01
## 12         Asheville 01-11-010 Municipality       19.88   Dec 2015-12-01
```

```
combined.data<- merge(ash, durham15, all.x = TRUE, all.y = TRUE)

ggplot(combined.data, aes(x= Date, y= Max.Day.Use, color = Water.System.Name)) +
  geom_point()+
  geom_line()+
labs(x = "Date",y = "Maximum Daily Withdrawals")
```

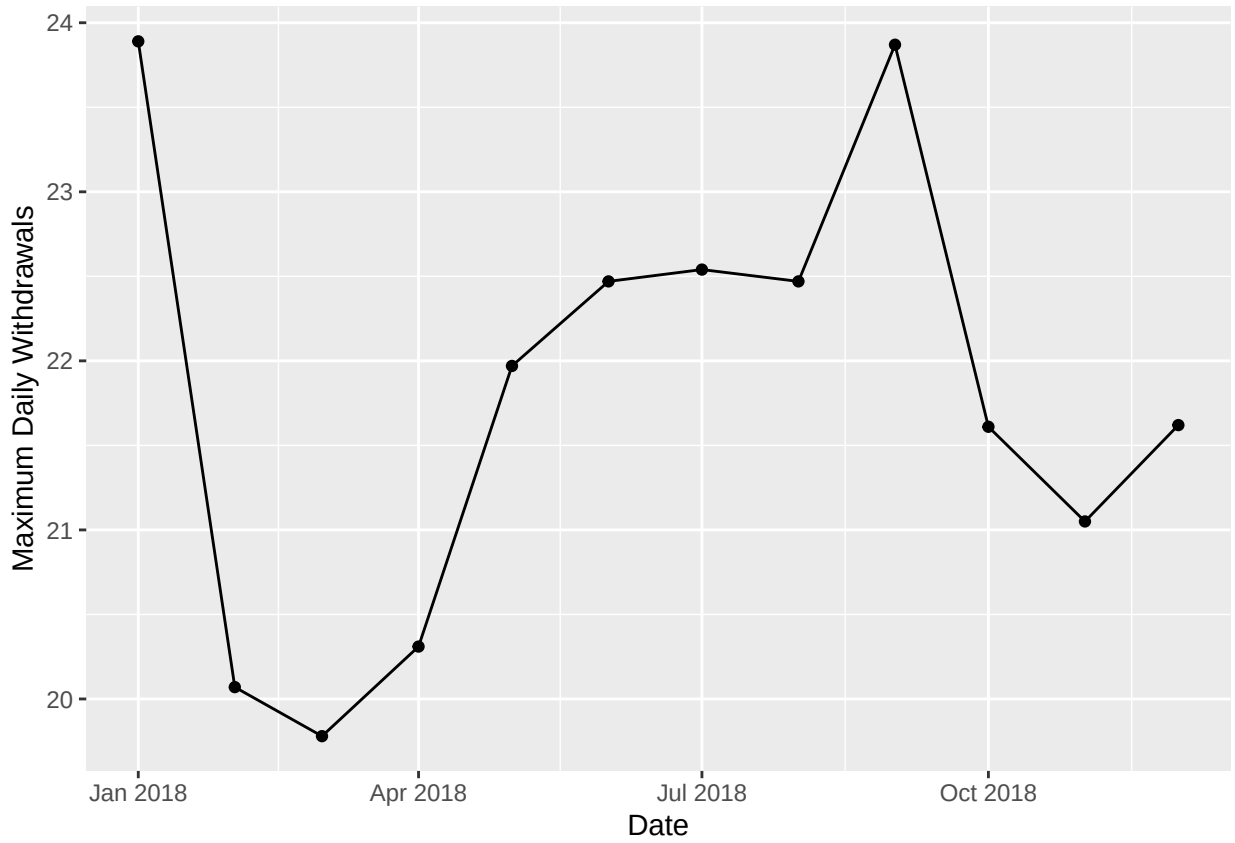


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2018 thru 2022.Add a smoothed line to the plot (method = 'loess').

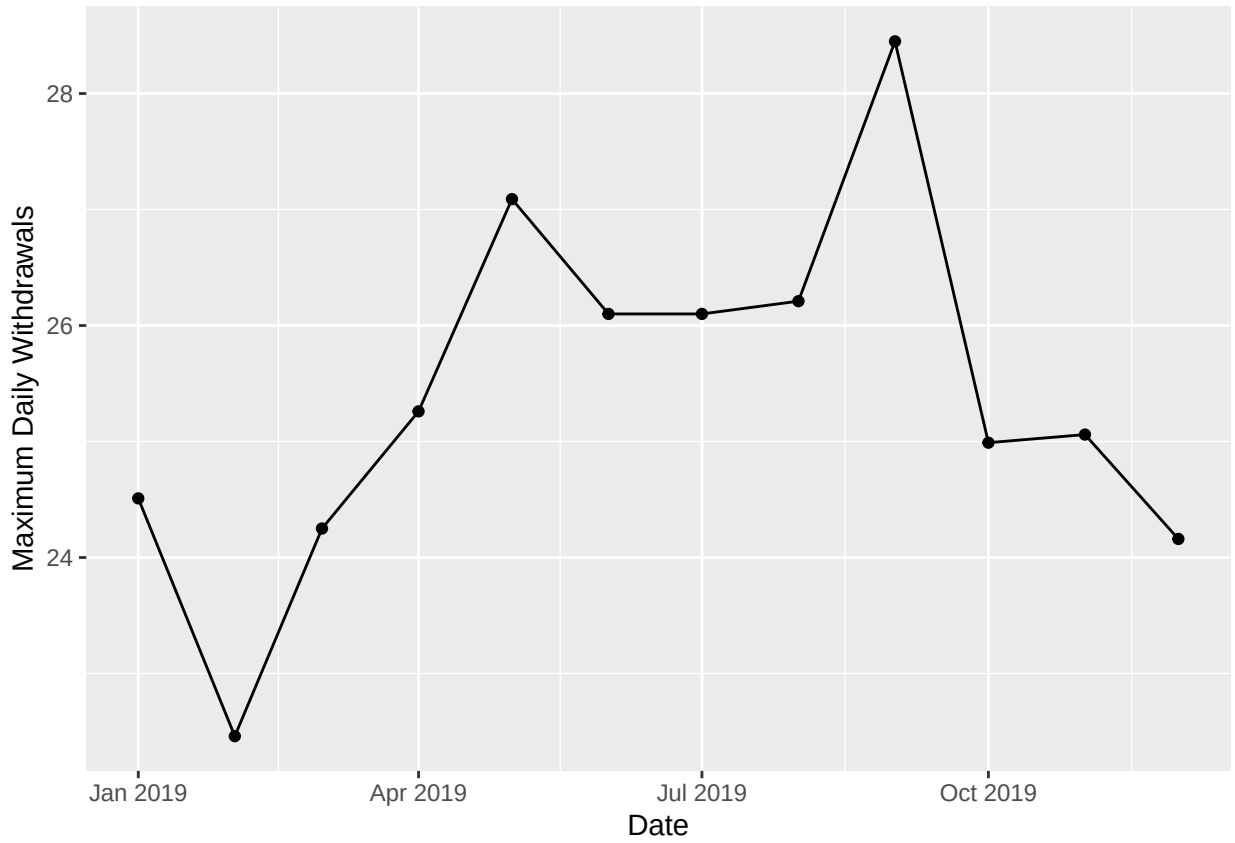
   TIP: See Section 3.2 in the "10_Data_Scraping.Rmd" where we apply "map2()" to iteratively run a function over two inputs. Pipe the output of the map2() function to `bindrows()` to combine the dataframes into a single one.

```
#9

the_years = seq(2018,2022)
the_site = rep('01-11-010', length(the_years))
ash.max.water <- map2(the_site, the_years, scrape.it) %>%
  bind_rows()
```

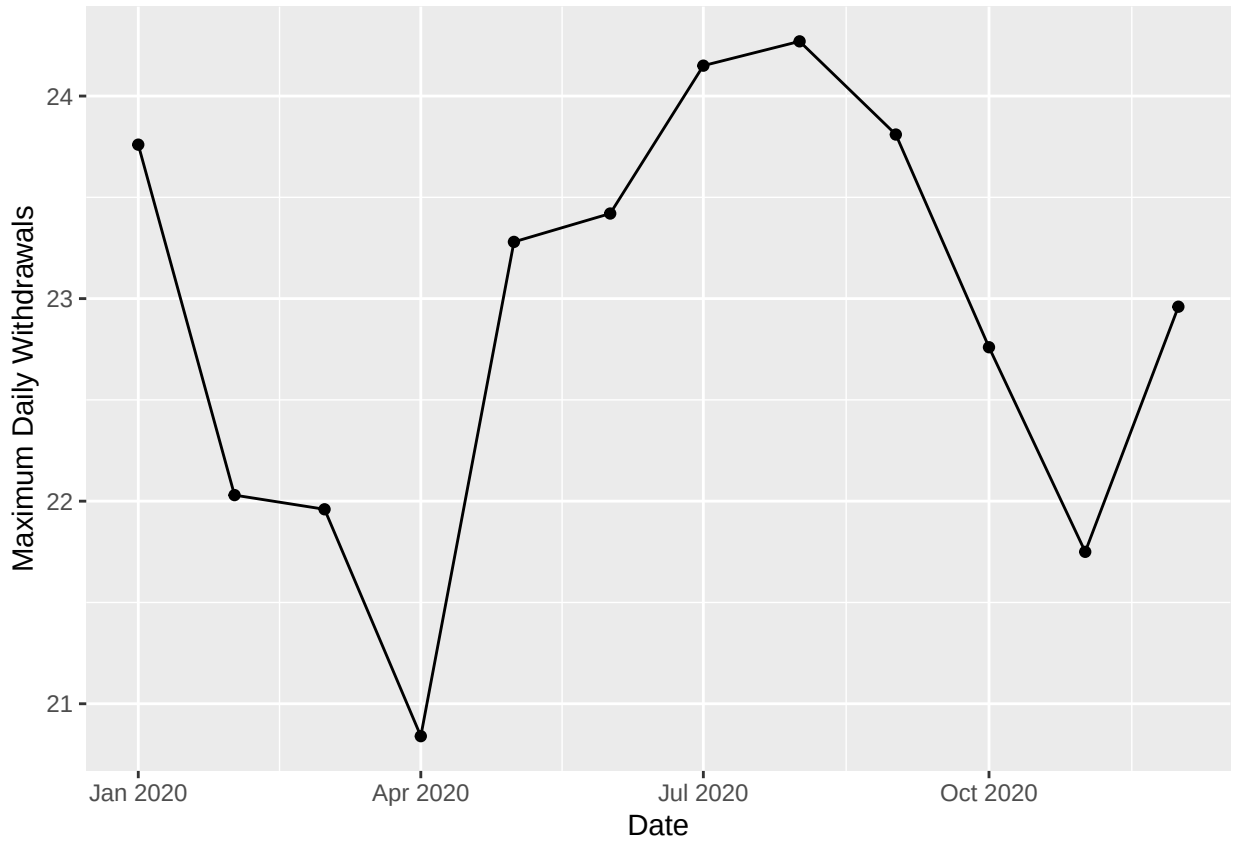
```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2018"
```



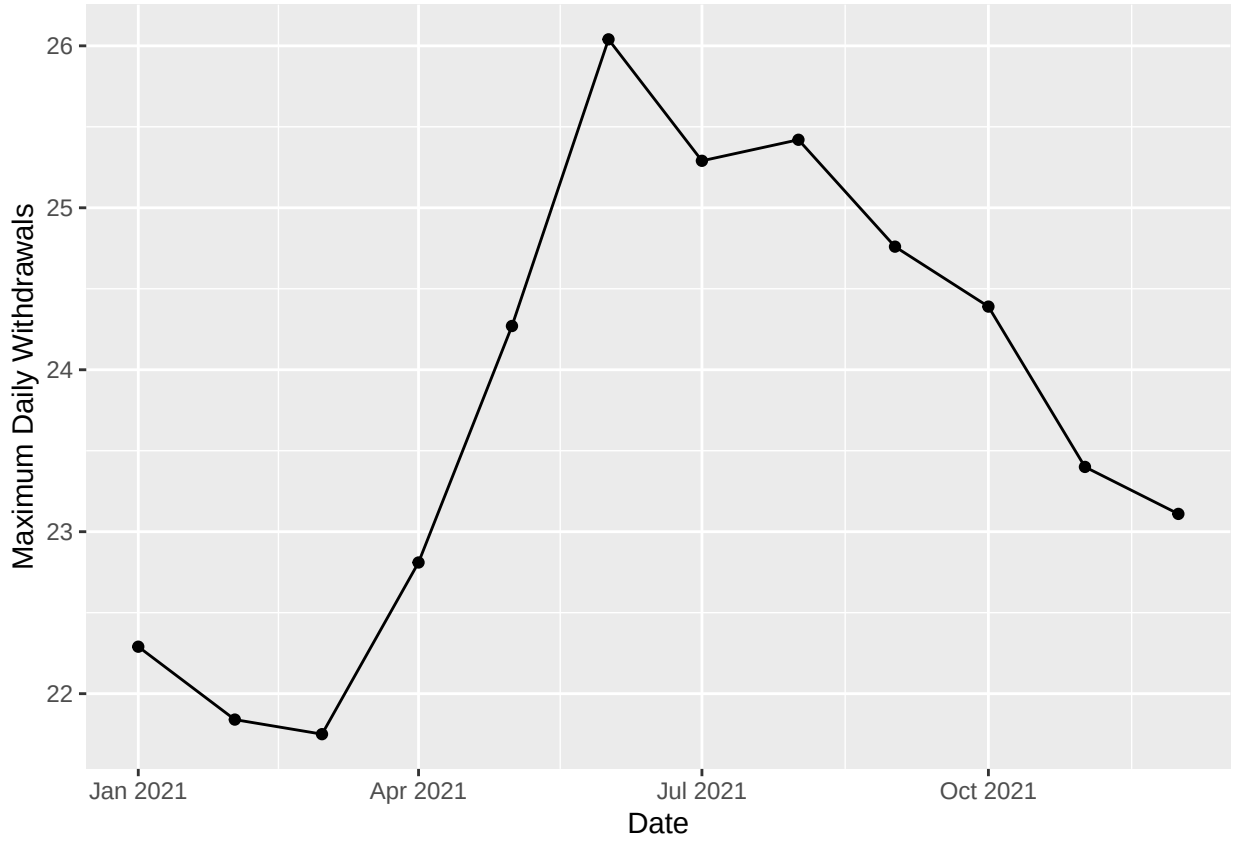

```
##    Water.System.Name     PWSID    Ownership Max.Day.Use Month       Date
## 1          Asheville 01-11-010 Municipality       23.89   Jan 2018-01-01
## 2          Asheville 01-11-010 Municipality       20.07   Feb 2018-02-01
## 3          Asheville 01-11-010 Municipality       19.78   Mar 2018-03-01
## 4          Asheville 01-11-010 Municipality       20.31   Apr 2018-04-01
## 5          Asheville 01-11-010 Municipality       21.97   May 2018-05-01
## 6          Asheville 01-11-010 Municipality       22.47   Jun 2018-06-01
## 7          Asheville 01-11-010 Municipality       22.54   Jul 2018-07-01
## 8          Asheville 01-11-010 Municipality       22.47   Aug 2018-08-01
## 9          Asheville 01-11-010 Municipality       23.87  Sept 2018-09-01
## 10         Asheville 01-11-010 Municipality       21.61   Oct 2018-10-01
## 11         Asheville 01-11-010 Municipality       21.05   Nov 2018-11-01
## 12         Asheville 01-11-010 Municipality       21.62   Dec 2018-12-01
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2019"
```



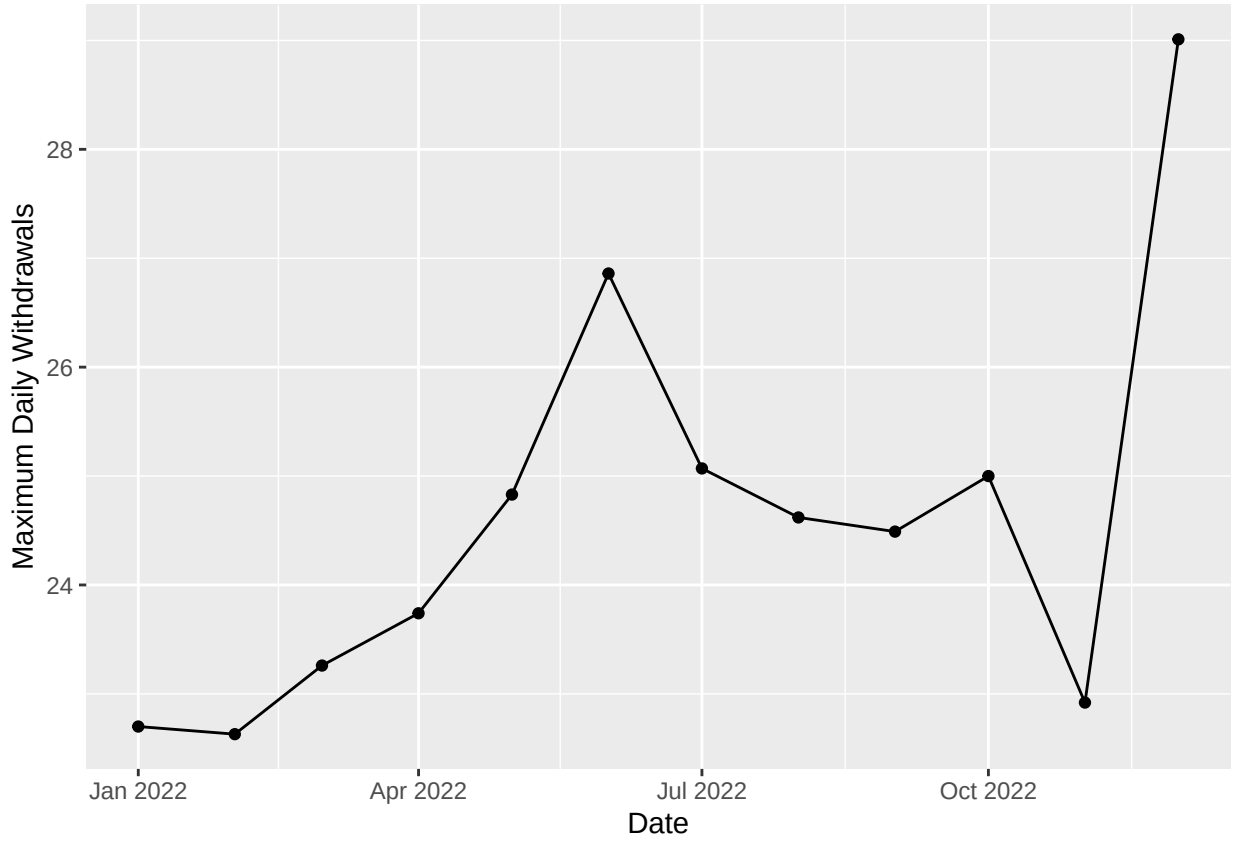

```
##    Water.System.Name     PWSID    Ownership Max.Day.Use Month       Date
## 1          Asheville 01-11-010 Municipality       24.51   Jan 2019-01-01
## 2          Asheville 01-11-010 Municipality       22.46   Feb 2019-02-01
## 3          Asheville 01-11-010 Municipality       24.25   Mar 2019-03-01
## 4          Asheville 01-11-010 Municipality       25.26   Apr 2019-04-01
## 5          Asheville 01-11-010 Municipality       27.09   May 2019-05-01
## 6          Asheville 01-11-010 Municipality       26.10   Jun 2019-06-01
## 7          Asheville 01-11-010 Municipality       26.10   Jul 2019-07-01
## 8          Asheville 01-11-010 Municipality       26.21   Aug 2019-08-01
## 9          Asheville 01-11-010 Municipality       28.45  Sept 2019-09-01
## 10         Asheville 01-11-010 Municipality       24.99   Oct 2019-10-01
## 11         Asheville 01-11-010 Municipality       25.06   Nov 2019-11-01
## 12         Asheville 01-11-010 Municipality       24.16   Dec 2019-12-01
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2020"
```




```
##    Water.System.Name     PWSID    Ownership Max.Day.Use Month       Date
## 1          Asheville 01-11-010 Municipality       23.76   Jan 2020-01-01
## 2          Asheville 01-11-010 Municipality       22.03   Feb 2020-02-01
## 3          Asheville 01-11-010 Municipality       21.96   Mar 2020-03-01
## 4          Asheville 01-11-010 Municipality       20.84   Apr 2020-04-01
## 5          Asheville 01-11-010 Municipality       23.28   May 2020-05-01
## 6          Asheville 01-11-010 Municipality       23.42   Jun 2020-06-01
## 7          Asheville 01-11-010 Municipality       24.15   Jul 2020-07-01
## 8          Asheville 01-11-010 Municipality       24.27   Aug 2020-08-01
## 9          Asheville 01-11-010 Municipality       23.81  Sept 2020-09-01
## 10         Asheville 01-11-010 Municipality       22.76   Oct 2020-10-01
## 11         Asheville 01-11-010 Municipality       21.75   Nov 2020-11-01
## 12         Asheville 01-11-010 Municipality       22.96   Dec 2020-12-01
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2021"
```



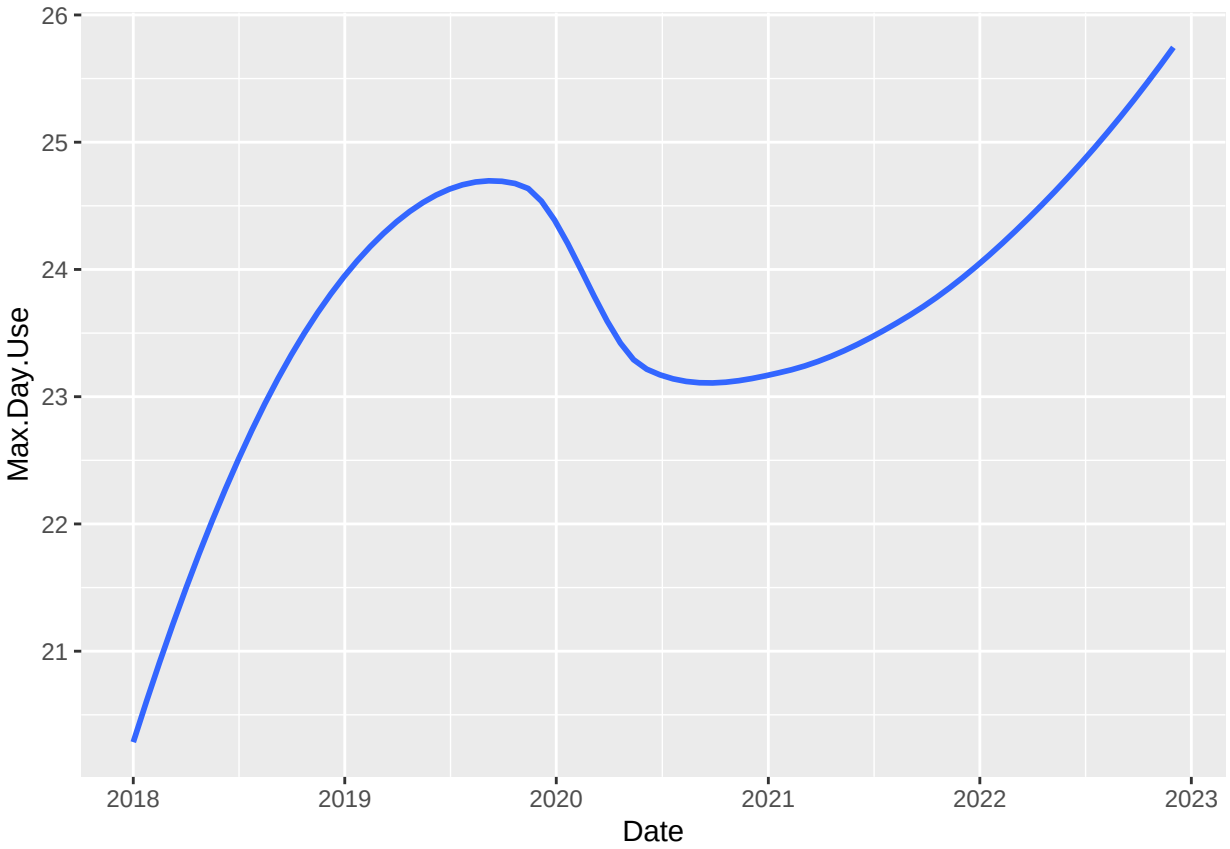

```
##    Water.System.Name     PWSID    Ownership Max.Day.Use Month       Date
## 1          Asheville 01-11-010 Municipality       22.29   Jan 2021-01-01
## 2          Asheville 01-11-010 Municipality       21.84   Feb 2021-02-01
## 3          Asheville 01-11-010 Municipality       21.75   Mar 2021-03-01
## 4          Asheville 01-11-010 Municipality       22.81   Apr 2021-04-01
## 5          Asheville 01-11-010 Municipality       24.27   May 2021-05-01
## 6          Asheville 01-11-010 Municipality       26.04   Jun 2021-06-01
## 7          Asheville 01-11-010 Municipality       25.29   Jul 2021-07-01
## 8          Asheville 01-11-010 Municipality       25.42   Aug 2021-08-01
## 9          Asheville 01-11-010 Municipality       24.76  Sept 2021-09-01
## 10         Asheville 01-11-010 Municipality       24.39   Oct 2021-10-01
## 11         Asheville 01-11-010 Municipality       23.40   Nov 2021-11-01
## 12         Asheville 01-11-010 Municipality       23.11   Dec 2021-12-01
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2022"
```




```
##    Water.System.Name     PWSID    Ownership Max.Day.Use Month       Date
## 1          Asheville 01-11-010 Municipality       22.70   Jan 2022-01-01
## 2          Asheville 01-11-010 Municipality       22.63   Feb 2022-02-01
## 3          Asheville 01-11-010 Municipality       23.26   Mar 2022-03-01
## 4          Asheville 01-11-010 Municipality       23.74   Apr 2022-04-01
## 5          Asheville 01-11-010 Municipality       24.83   May 2022-05-01
## 6          Asheville 01-11-010 Municipality       26.86   Jun 2022-06-01
## 7          Asheville 01-11-010 Municipality       25.07   Jul 2022-07-01
## 8          Asheville 01-11-010 Municipality       24.62   Aug 2022-08-01
## 9          Asheville 01-11-010 Municipality       24.49  Sept 2022-09-01
## 10         Asheville 01-11-010 Municipality       25.00   Oct 2022-10-01
## 11         Asheville 01-11-010 Municipality       22.92   Nov 2022-11-01
## 12         Asheville 01-11-010 Municipality       29.01   Dec 2022-12-01
```

```
ash.max.water %>%
ggplot(aes(x= Date, y= Max.Day.Use)) +
#geom_line() +
geom_smooth(method='loess',se=FALSE) +
scale_x_date(date_breaks = '1 year', date_labels = '%Y')
```

```
## `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 asheville has a trend indicating increased water usage over time. > There is a dip between 2020 and 2022 during covid times.