Assignment 10: Data Scraping

Robert Hill

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(lubridate);library(here);library(scales)
getwd()
```

```
## [1] "/home/guest/EDA-Spring2023-RH"
```

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

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 <- html_node(theURL,"div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()

PWSID <- html_node(theURL,"td tr:nth-child(1) td:nth-child(5)") %>%
  html_text()

ownership <- html_node(theURL,"div+ table tr:nth-child(2) td:nth-child(4)") %>%
  html_text()

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

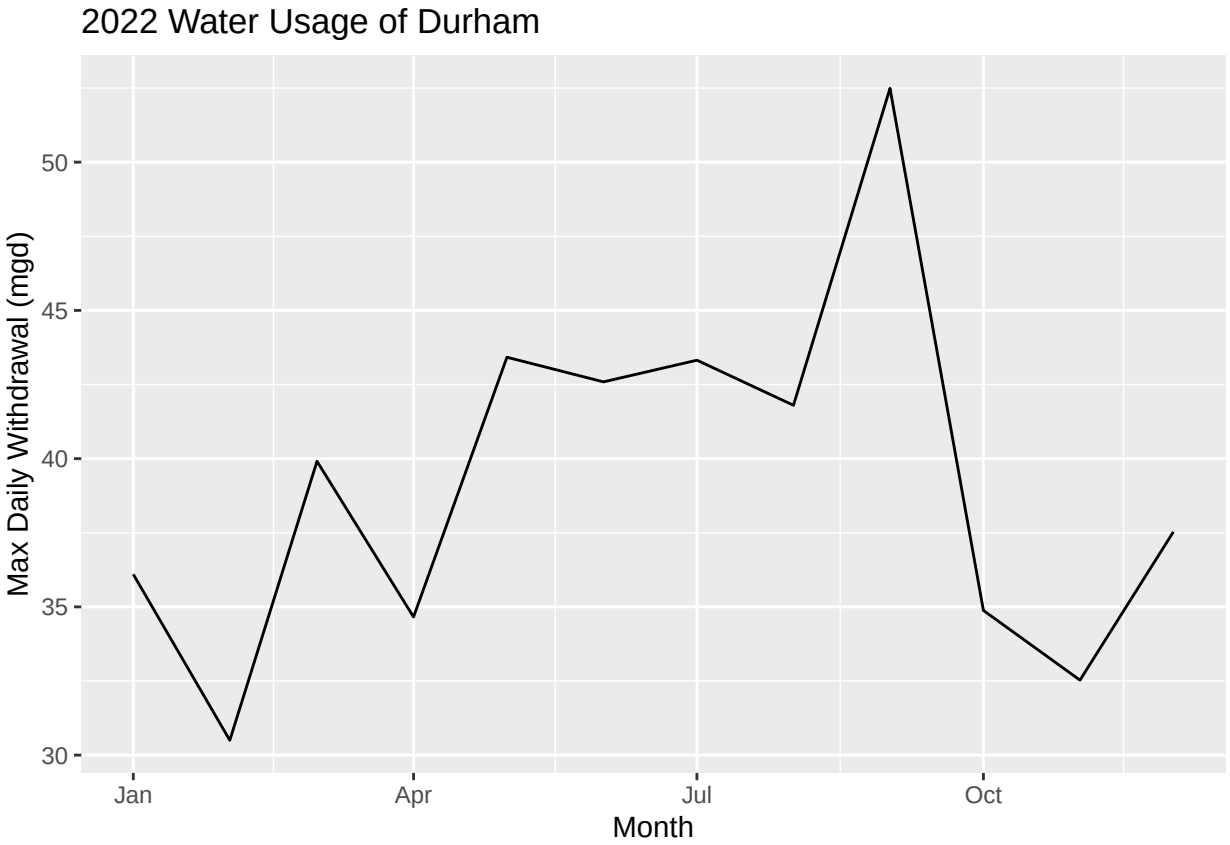
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   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 max daily withdrawals across the months for 2022

```
#4
the_df <- data.frame(
  "Month" = c(1,5,9,2,6,10,3,7,11,4,8,12),
  "Water_System_Name" = water.system.name,
  "PWSID" = PWSID,
  "Ownership" = ownership,
  "Max_Withdrawals_mgd" = as.numeric(max.withdrawals.mgd)
)
the_df$Date <- (my(paste(the_df$Month,"-2022")))


#5
ggplot(the_df,aes(x=Date,y=Max_Withdrawals_mgd)) +
  geom_line() +
  labs(title = paste("2022 Water Usage of Durham"),
       y="Max Daily Withdrawal (mgd)",
       x="Month") +
  scale_x_date(date_labels = "%b")
```



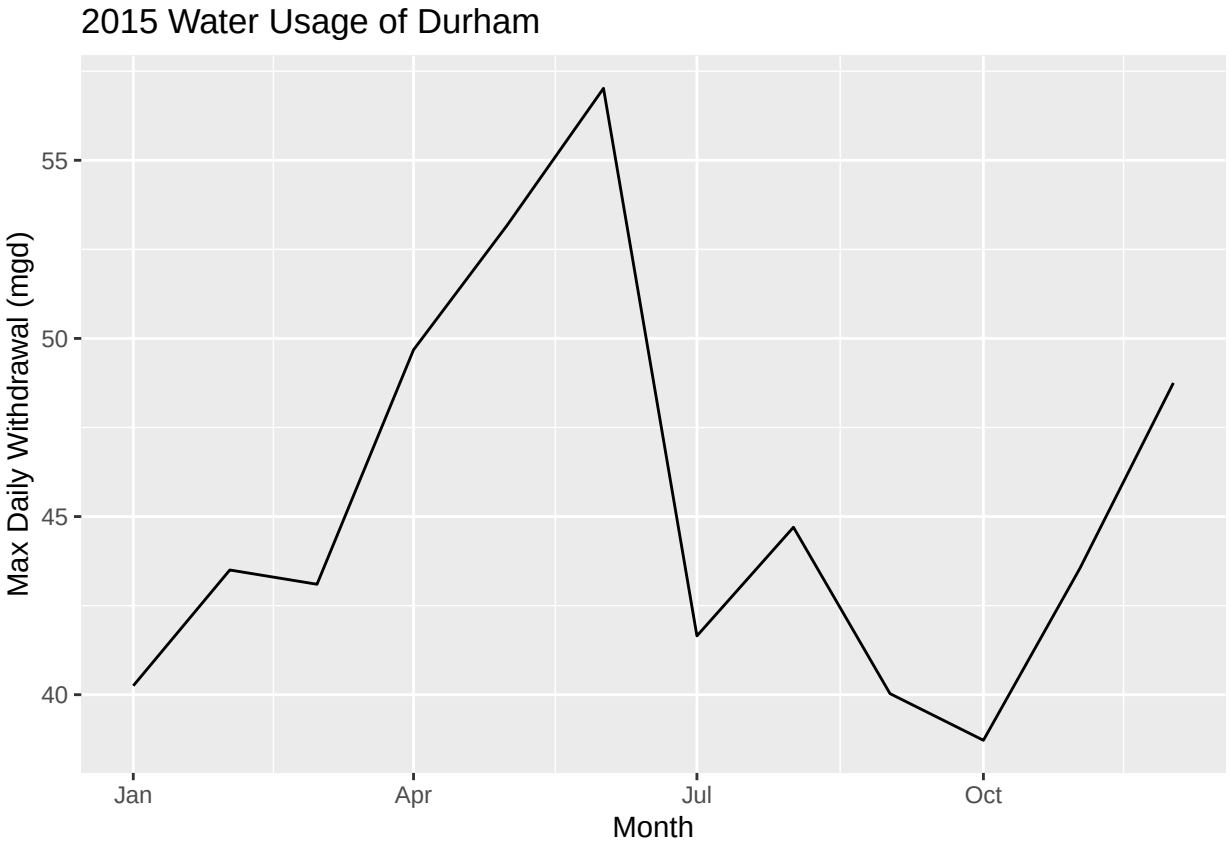
6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

```
#6.
#function with plot
scrape.it <- function(PWSID,the_year){
  the_URL <- read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',
                              PWSID,'&year=',the_year))


  water.system.name <- html_node(the_URL,"div+ table tr:nth-child(1) td:nth-child(2)") %>%
    html_text()
  PWSID <- html_node(the_URL,"td tr:nth-child(1) td:nth-child(5)") %>%
    html_text()
  ownership <- html_node(the_URL,"div+ table tr:nth-child(2) td:nth-child(4)") %>%
    html_text()
  max.withdrawals.mgd <- html_nodes(the_URL,"th~ td+ td") %>%
    html_text()
  df_function <- data.frame(
    "Month" = c(1,5,9,2,6,10,3,7,11,4,8,12),
    "Water_System_Name" = water.system.name,
    "PWSID" = PWSID,
    "Ownership" = ownership,
    "Max_Withdrawals_mgd" = as.numeric(max.withdrawals.mgd)
    )
  df_function$Date <- (my(paste(the_df$Month,"-",the_year)))
```

```
  plot_function <- ggplot(df_function,aes(x=Date,y=Max_Withdrawals_mgd)) +
    geom_line() +
    labs(title = paste(the_year, "Water Usage of",water.system.name),
         y="Max Daily Withdrawal (mgd)", x="Month",) +
    scale_x_date(date_labels = "%b")


  print(plot_function)
  return(df_function)
}

#function without plot
scrape.it2 <- function(PWSID,the_year){
  the_URL <- read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',
                              PWSID,'&year=',the_year))


  water.system.name <- html_node(the_URL,"div+ table tr:nth-child(1) td:nth-child(2)") %>%
    html_text()
  PWSID <- html_node(the_URL,"td tr:nth-child(1) td:nth-child(5)") %>%
    html_text()
  ownership <- html_node(the_URL,"div+ table tr:nth-child(2) td:nth-child(4)") %>%
    html_text()
  max.withdrawals.mgd <- html_nodes(the_URL,"th~ td+ td") %>%
    html_text()
  df_function <- data.frame(
    "Month" = c(1,5,9,2,6,10,3,7,11,4,8,12),
    "Water_System_Name" = water.system.name,
    "PWSID" = PWSID,
    "Ownership" = ownership,
    "Max_Withdrawals_mgd" = as.numeric(max.withdrawals.mgd)
    )
  df_function$Date <- (my(paste(the_df$Month,"-",the_year)))

  return(df_function)
}
```

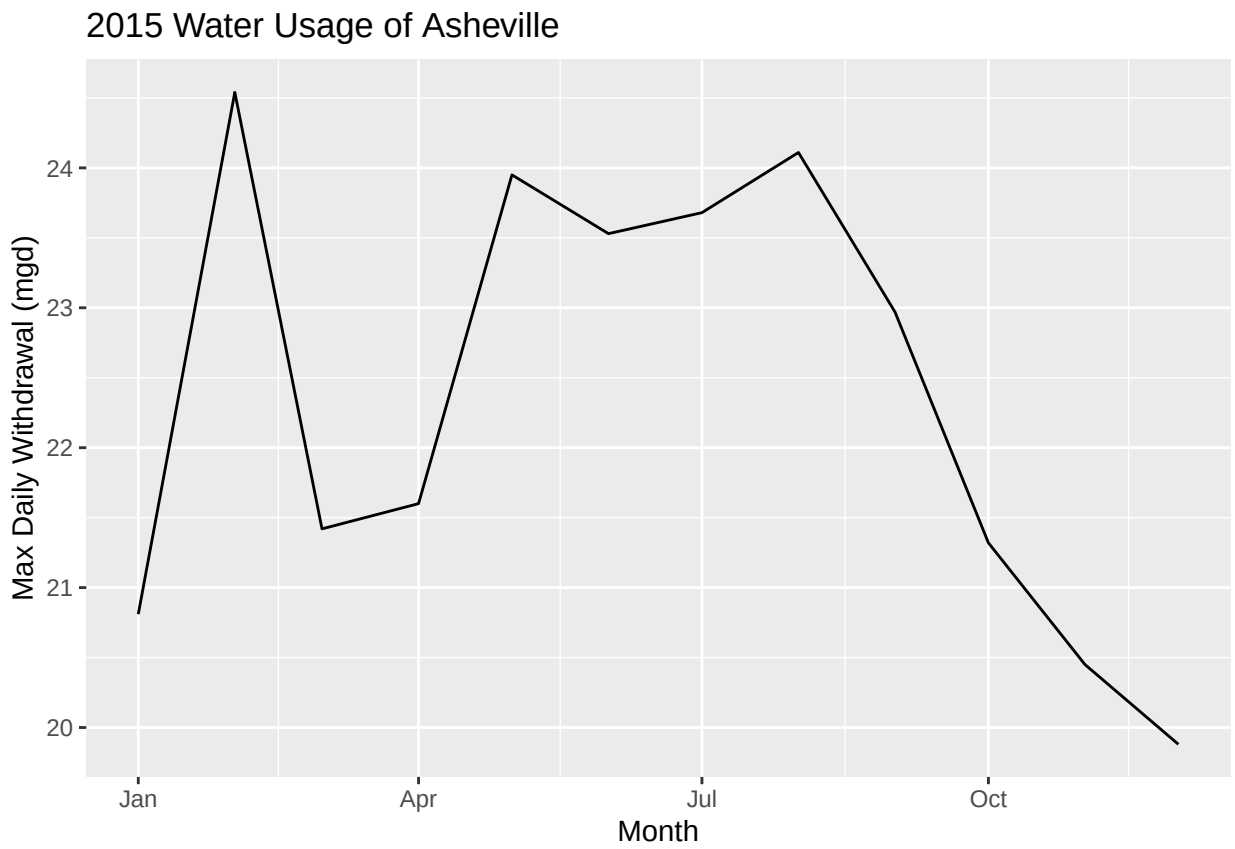
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID='03-32-010') for each month in 2015

```
#7
Durham2015 <-scrape.it('03-32-010',2015)
```

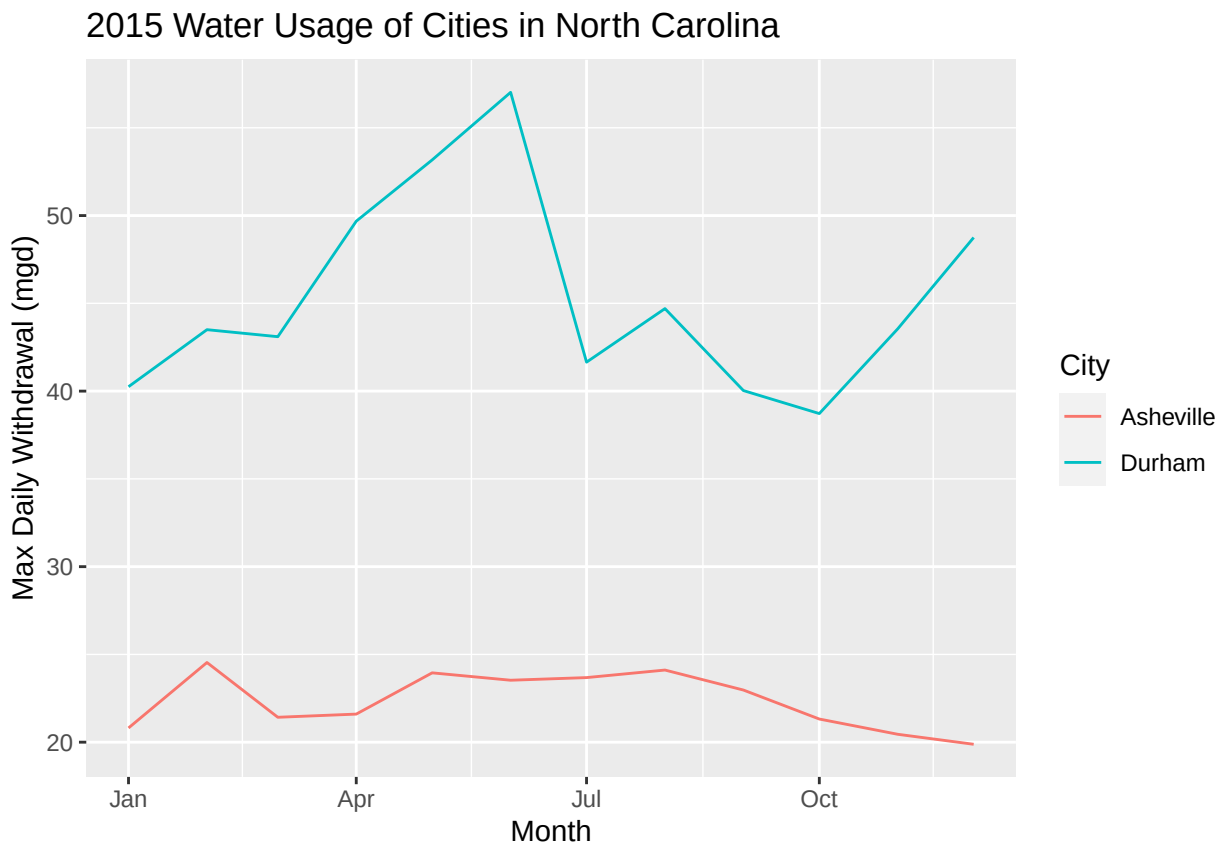


8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares Asheville's to Durham's water withdrawals.

```
#8
Asheville2015 <- scrape.it('01-11-010',2015)
```




```
combined2015 <- bind_rows(Asheville2015,Durham2015)

ggplot(combined2015,aes(x=Date,y=Max_Withdrawals_mgd,color=Water_System_Name)) +
  geom_line() +
  labs(title = paste("2015 Water Usage of Cities in North Carolina"),
       y="Max Daily Withdrawal (mgd)", x="Month", color="City") +
  scale_x_date(date_labels = "%b")
```

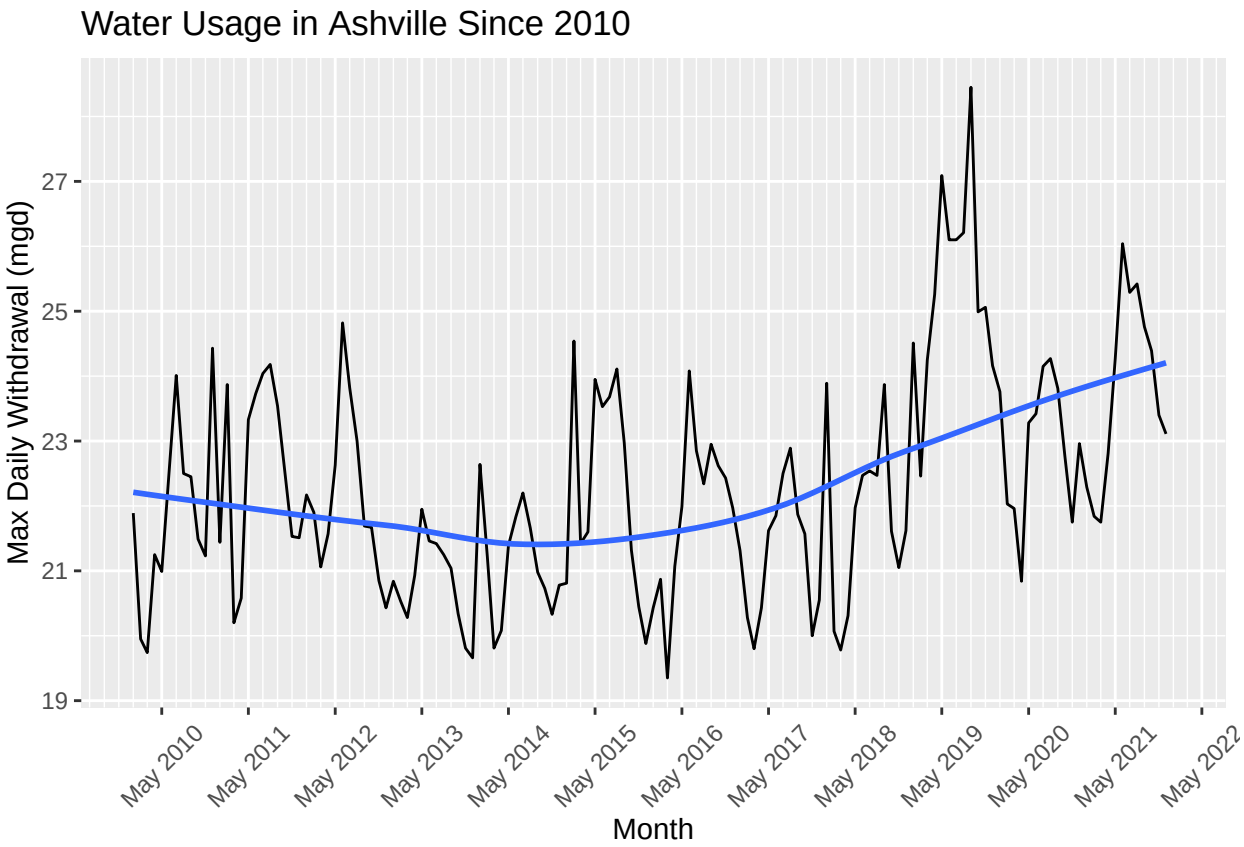


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2021.Add a smoothed line to the plot (method = 'loess').

   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.

```
#9
ID <- rep('01-11-010',12)
the_years <- seq(2010,2021)
df_Ashville <- map2(ID,the_years,scrape.it2) %>%
  bind_rows()

ggplot(df_Ashville,aes(x=Date,y=Max_Withdrawals_mgd)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("Water Usage in Ashville Since 2010"),
       y="Max Daily Withdrawal (mgd)",
       x="Month") +
  scale_x_date(date_labels = "%b %Y", date_breaks = "12 months",
               limits=as.Date(c("2010-01-01","2022-01-01")),
               minor_breaks = "2 months") +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5))
```

```
## `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, there is an increasing trend in water usage.