Assignment 10: Data Scraping

Griffin Bird

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(rvest)
```

```
## Warning: package 'rvest' was built under R version 4.2.3
```

```
library(here)
library(tidyverse, quietly = TRUE)
library(lubridate)
library(ggplot2)
library(ggthemes)
library(dplyr)

mytheme <- theme_wsj() + theme(plot.title = element_text(hjust = 0.5),
   panel.grid.minor = element_line(color = 2,
   size = 0.25,
   linetype = 1),
  legend.position = "bottom",
  axis.text = element_text(face="plain", size = 12),
  axis.title=element_text(size=12),
  title = element_text(size=12))
```

```
## Warning: The ‘size‘ argument of ‘element_line()‘ is deprecated as of ggplot2 3.4.0.
## i Please use the ‘linewidth‘ argument instead.
```

```
theme_set(mytheme)

here
```

```
## function (...)
## {
##     .root_env$root$f(...)
## }
## <bytecode: 0x000001d3ac48b958>
## <environment: namespace:here>
```

```
getwd()
```

```
## [1] "C:/Users/griff/OneDrive/Desktop/EDA-Spring2023/Assignments"
```

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
webpage <- 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 "27.6400".

```
#3
water.system.name <- webpage %>%
  html_node("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()
water.system.name
```

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

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

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

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

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

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

```
##  [1] "36.1000" "43.4200" "52.4900" "30.5000" "42.5900" "34.8800" "39.9100"
##  [8] "43.3200" "32.5300" "34.6600" "41.8000" "37.5300"
```

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.)

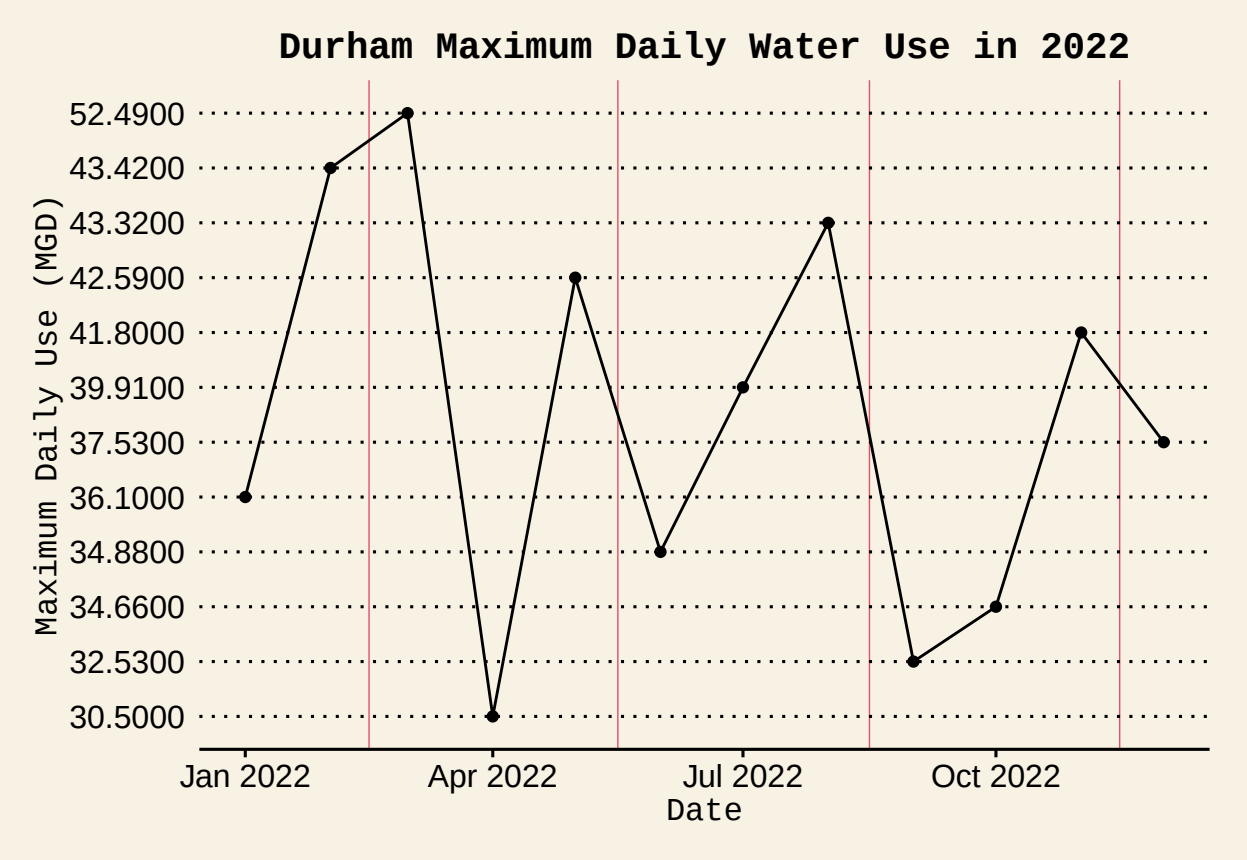
   TIP: Use `rep()` to repeat a value when creating a dataframe.

   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 average daily withdrawals across the months for 2022

```
#4
WithdrawalData <- data.frame("Water System Name"=rep(c(water.system.name)),
                             "PWSID"=rep(c(PWSID)),
                             "Ownership"=rep(c(ownership)),
                             "Max Daily Use"=c(max.withdrawals.mgd),
```

```
                          "Month" = c(month.abb),
                          Year = rep(c(2022)))

WithdrawalData <- WithdrawalData %>%
  mutate(Date = my(paste(Month,"-",Year))) %>%
  select(Water.System.Name, PWSID, Ownership, Max.Daily.Use, Month, Date)

#5

MaxUsePlot <- ggplot(WithdrawalData, aes(x = Date, y = max.withdrawals.mgd, group=1)) +
  labs(title = "Durham Maximum Daily Water Use in 2022",
       x = "Date",
       y = "Maximum Daily Use (MGD)")  +
  geom_line() +
  geom_point()

MaxUsePlot
```



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**.

```
the_base_url <- 'https://www.ncwater.org/WUDC/app/LWSP/'
the_facility <- '03-32-010'
the_year <- 2015
```

```
the_scrape_url <- paste0(the_base_url,'report.php?pwsid=', the_facility,'&year=', the_year)

the_website <- read_html(the_scrape_url)


water.system.name_tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'
PWSID_tag <- 'td tr:nth-child(1) td:nth-child(5)'
ownership_tag <- 'div+ table tr:nth-child(2) td:nth-child(4)'
max.withdrawals_tag <- 'th~ td+ td'

the_water.system <- the_website %>% html_nodes(water.system.name_tag) %>% html_text()
the_PWSID <- the_website %>%  html_node(PWSID_tag) %>% html_text()
the_ownership.name <- the_website %>% html_node(ownership_tag) %>% html_text()
the_max.withdrawals.mgd <- the_website %>% html_nodes(max.withdrawals_tag) %>% html_text()


ScrapedData_df <- data.frame("Water System Name"=rep(c(the_water.system)),
                             "PWSID"=rep(c(the_PWSID)),
                             "Ownership"=rep(c(the_ownership.name)),
                             "Max Daily Use"=c(as.numeric(the_max.withdrawals.mgd)),
                             "Month" = c(month.abb),
                             Year = rep(c(the_year)))%>%
  mutate(Date = my(paste(Month,"-",Year))) %>%
  select(Water.System.Name, PWSID, Ownership, Max.Daily.Use, Month, Date)
```

Run the above; it should produce the same result as the previous R chunk. HOWEVER, change the year variable to `2015` and re-run the chunk. Change the facility ID to `0218-0238` and run again. Now, we have a nifty little scraping tool!

2.2 Automation, Step 1: Build a function

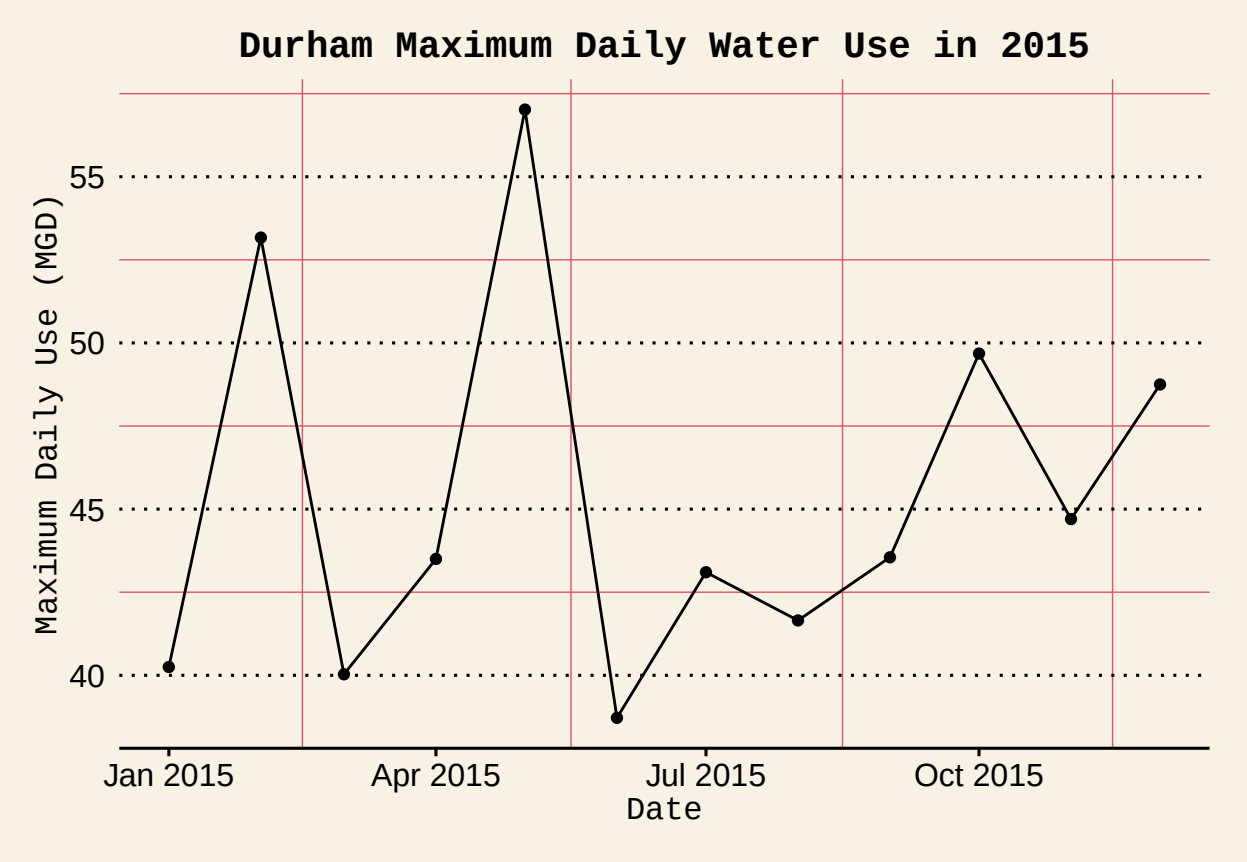
We have our code so we can fairly easily scrape any site (if we know its ID) for any year. Let's improve our code so we can automate the process more easily and perhaps scrape many years worth of data. To make this process run more easily, we'll first convert our code into a function that produces a dataframe of withdrawal data for a given year and facility ID.

```
scrape.it <- function(the_year, the_facility){


the_website <- read_html(paste0
('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=', the_facility,'&year=', the_year))

water.system.name_tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'
PWSID_tag <- 'td tr:nth-child(1) td:nth-child(5)'
ownership_tag <- 'div+ table tr:nth-child(2) td:nth-child(4)'
max.withdrawals_tag <- 'th~ td+ td'

the_water.system <- the_website %>% html_nodes(water.system.name_tag) %>% html_text()
the_PWSID <- the_website %>%  html_node(PWSID_tag) %>% html_text()
the_ownership.name <- the_website %>% html_node(ownership_tag) %>% html_text()
the_max.withdrawals.mgd <- the_website %>% html_nodes(max.withdrawals_tag) %>% html_text()
```

```
ScrapedData_df <- data.frame("Water System Name"=rep(c(the_water.system)),
                             "PWSID"=rep(c(the_PWSID)),
                             "Ownership"=rep(c(the_ownership.name)),
                             "Max Daily Use"=c(as.numeric(the_max.withdrawals.mgd)),
                             "Month" = c(month.abb),
                             Year = rep(c(the_year))) %>%
  mutate(Date = my(paste(Month,"-",Year))) %>%
  select(Water.System.Name, PWSID, Ownership, Max.Daily.Use, Month, Date)

  return(ScrapedData_df)
}
```

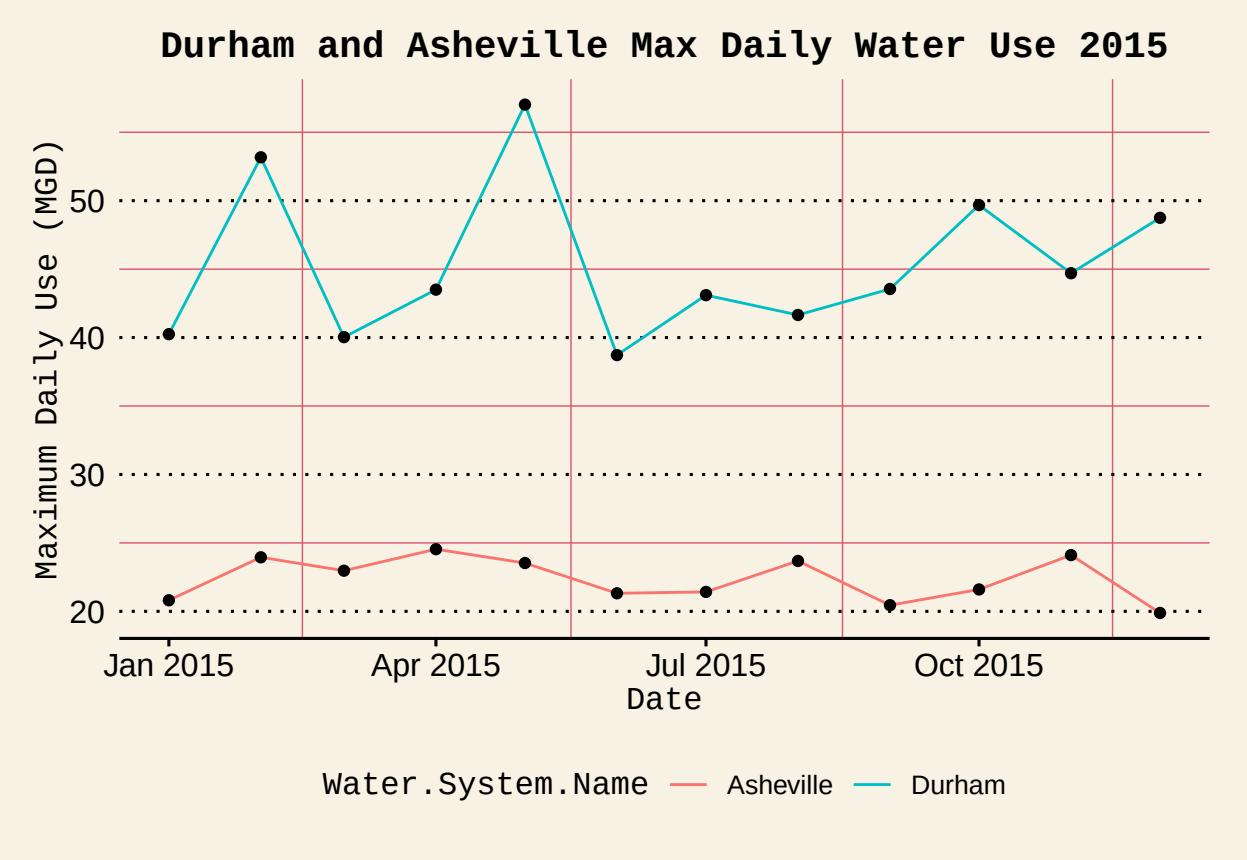
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID=‘03-32-010’) for each month in 2015

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

Durham2015.plot <- ggplot(Durham2015, aes(x = Date, y = Max.Daily.Use, group=1)) +
  labs(title = "Durham Maximum Daily Water Use in 2015",
       x = "Date",
       y = "Maximum Daily Use (MGD)")  +
  geom_line() +
  geom_point()

print(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

Asheville2015 <- scrape.it(2015, '01-11-010')
view(Asheville2015)

DurAsh <- rbind(Durham2015, Asheville2015)

DurhamvAsheville <- ggplot(DurAsh, aes(x = Date, y = Max.Daily.Use)) +
  labs(title = "Durham and Asheville Max Daily Water Use 2015",
       x = "Date",
       y = "Maximum Daily Use (MGD)")  +
  geom_line(aes(color= Water.System.Name))+
  geom_point()

print(DurhamvAsheville)
```



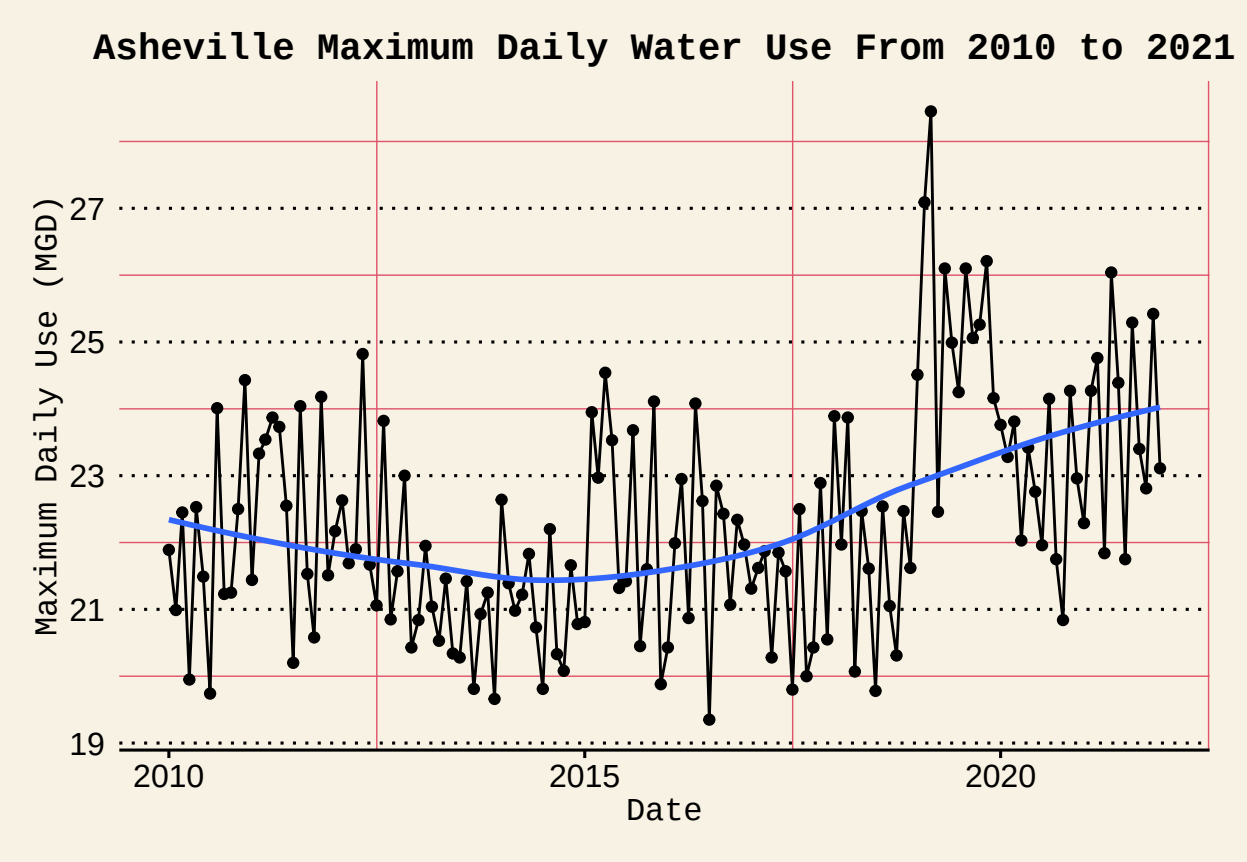
9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2021.Add a smoothed line to the plot (method = 'loess').

   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

the_years = rep(2010:2021)
my_facility = '01-11-010'

the_dfs <- lapply(X = the_years,
                  FUN = scrape.it,
                  the_facility = my_facility)

the_Asheville_df <- bind_rows(the_dfs)

Durham2010to2021 <- ggplot(the_Asheville_df, aes(x = Date, y = Max.Daily.Use)) +
  labs(title = "Asheville Maximum Daily Water Use From 2010 to 2021",
       x = "Date",
       y = "Maximum Daily Use (MGD)")  +
  geom_line() +
  geom_point()+
  geom_smooth(method="loess",se=FALSE)
```

```
print(Durham2010to2021)
```

```
## `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?

There looks to be an upward trend in Asheville water use from 2010 to 2021, of course we could do some time series trend analysis to investigate further.