Assignment 10: Data Scraping

Jiahuan Li

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

```r
#1
library(tidyverse)
library(rvest)
library(lubridate)
library(xml2)
library(ggplot2)

library(here)
here()
```

```
## [1] "D:/Users/Lijh/Desktop/872 R & data analytics/ENVIR872"
```

```r
# Set theme
mytheme <- theme_classic() +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
theme_set(mytheme)
```

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
url <- "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2022"
webpage <- read_html(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), with the first value being "27.6400".

```
#3
system_info_table <- webpage %>%
  html_nodes("table:nth-child(7) td") %>%
  html_text()
system_info_table
```

```
##  [1] "Water System Name:"
##  [2] "Durham"
##  [3] " "
##  [4] "PWSID:"
##  [5] "03-32-010"
##  [6] "Mailing Address:"
##  [7] "101 City Hall PlazaDurham, NC 27701"
##  [8] "Ownership:"
##  [9] "Municipality"
## [10] " "
## [11] "Contact Person:"
## [12] "Sydney Miller"
## [13] "Title:"
## [14] "Water Resources Manager"
## [15] "Phone:"
## [16] "919-560-4381"
## [17] "Cell/Mobile:"
## [18] "--"
## [19] " "
## [20] "Secondary Contact:"
## [21] "Mary Tiger, Asst. Dir."
## [22] " "
## [23] "Phone:"
## [24] "919-560-4381"
## [25] "Mailing Address:"
## [26] "1600 Mist Lake DriveDurham, NC 27704"
## [27] "Cell/Mobile:"
## [28] "--"
```

```
water.system.name <- webpage %>%
  html_nodes("table:nth-child(7) 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("table:nth-child(7) 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.)

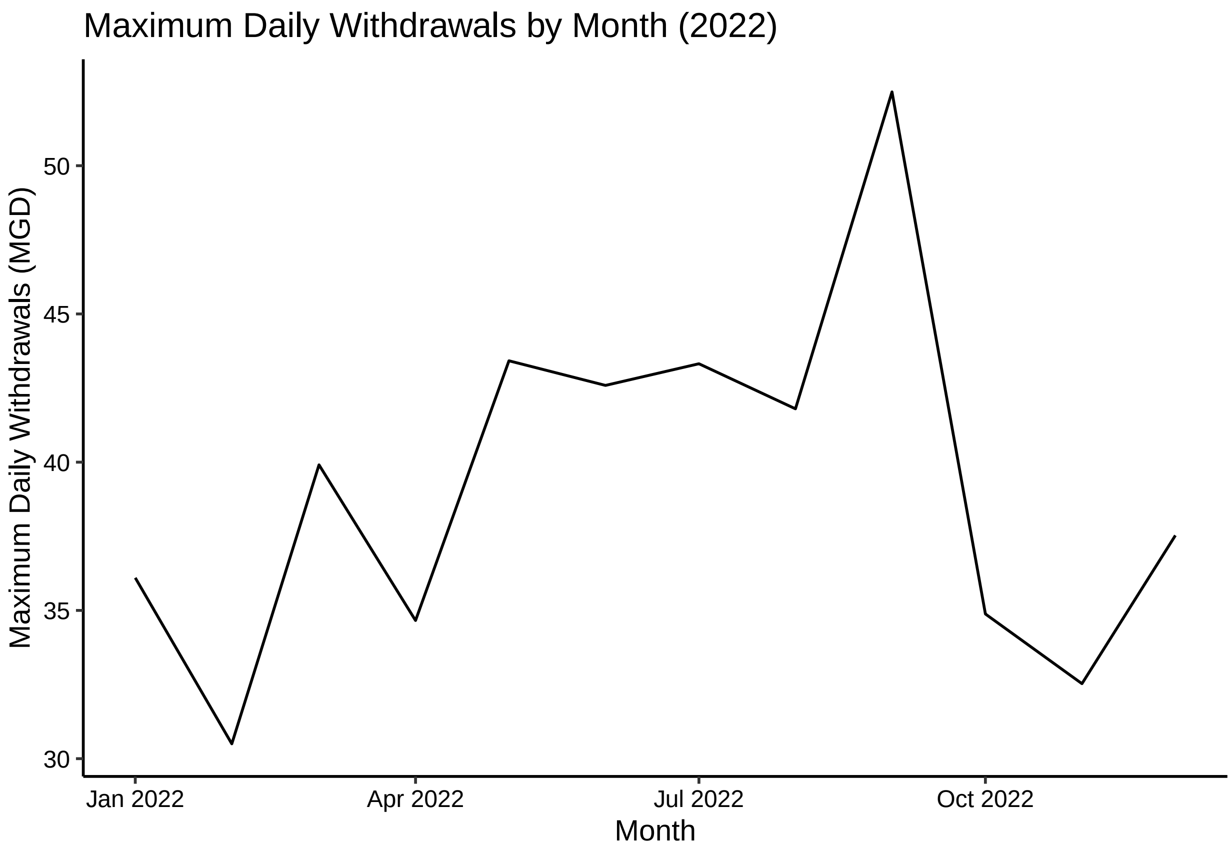
   TIP: Use `rep()` to repeat a value when creating a dataframe.

   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 max daily withdrawals across the months for 2022

```
#4
df <- data.frame(
  "water.system.name" = rep(water.system.name, 12),
  "PWSID" = rep(PWSID, 12),
  "ownership" = rep(ownership, 12),
  "Month" = c("Jan", "May", "Sep", "Feb", "Jun", "Oct", "Mar", "Jul", "Nov", "Apr", "Aug", "Dec"),
  "max.day.use" = as.numeric(max.withdrawals.mgd)
)

df$Date <- my(paste(df$Month,"-",2022))



#5
df <- arrange(df, Date)
```

```r
ggplot(df, aes(x = Date, y = max.day.use)) +
  geom_line() +
  labs(title = "Maximum Daily Withdrawals by Month (2022)",
       x = "Month",
       y = "Maximum Daily Withdrawals (MGD)")
```

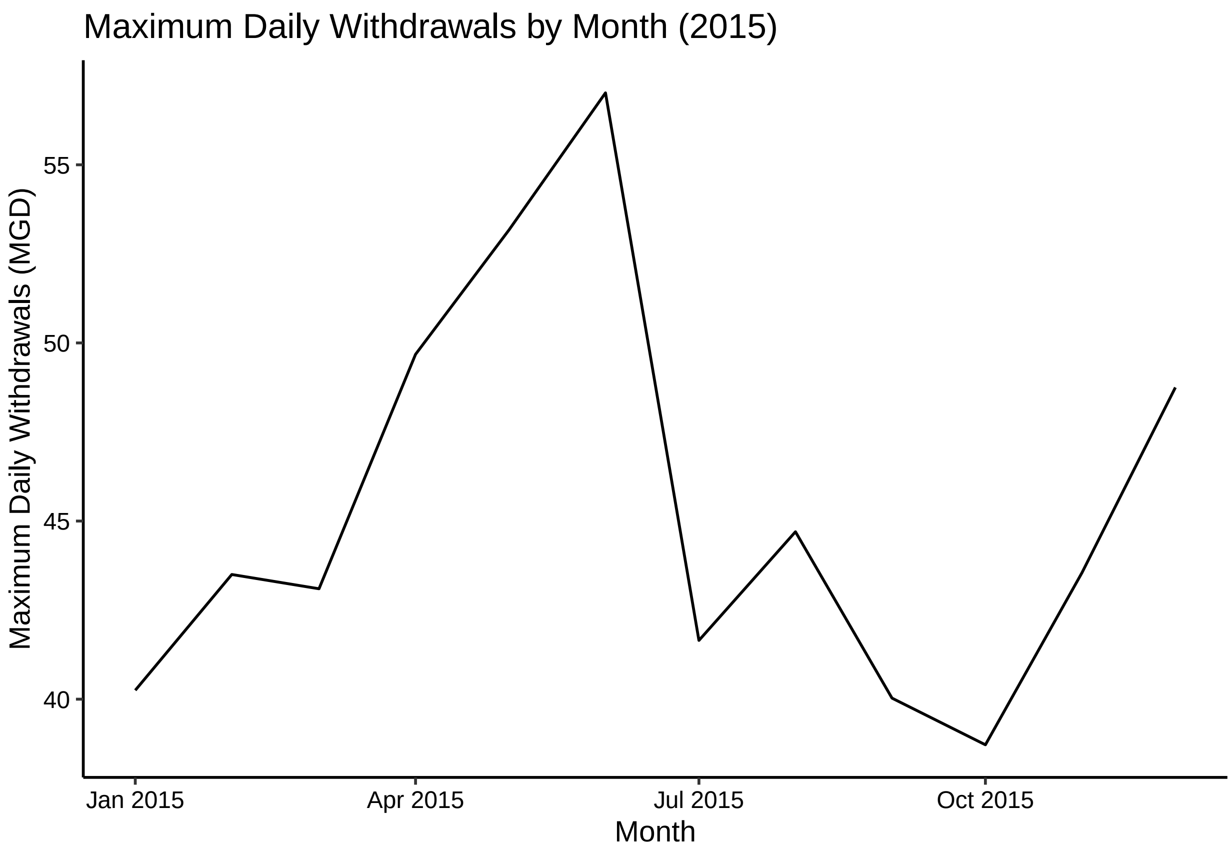


6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

```r
#6.
fun <- function(the_year, the_PWSID){

  #Retrieve the website contents
  url1 <- paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',the_PWSID,'&year=',the_year)
  the_website <- read_html(url1)

  water.system.name1 <- "table:nth-child(7) tr:nth-child(1) td:nth-child(2)"
  ownership1 <- "table:nth-child(7) tr:nth-child(2) td:nth-child(4)"
  max.withdrawals.mgd1 <- "th~ td+ td"

  #Scrape the data items
  the_system_name <- the_website %>% html_nodes(water.system.name1) %>% html_text()
  the_ownership <- the_website %>%   html_nodes(ownership1) %>%  html_text()
  the_mgd <- the_website %>% html_nodes(max.withdrawals.mgd1) %>% html_text()
  df_mgd <- data.frame(
  "water.system.name" = rep(the_system_name, 12),
  "PWSID" = rep(the_PWSID, 12),
  "Year" = rep(the_year, 12),
  "ownership" = rep(the_ownership, 12),
  "Month" = c("Jan", "May", "Sep", "Feb", "Jun", "Oct", "Mar", "Jul", "Nov", "Apr", "Aug", "Dec"),
  "max.day.use" = as.numeric(the_mgd)
)

  df_mgd <- df_mgd %>%
    mutate(Date = my(paste(Month,"-",the_year))) %>%
    arrange(df_mgd, Date)

  #Return the dataframe
  return(df_mgd)
}
```

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

```
#7
df_2015 <- fun(2015, '03-32-010')

ggplot(df_2015, aes(x = Date, y = max.day.use)) +
  geom_line() +
  labs(title = "Maximum Daily Withdrawals by Month (2015)",
       x = "Month",
       y = "Maximum Daily Withdrawals (MGD)")
```

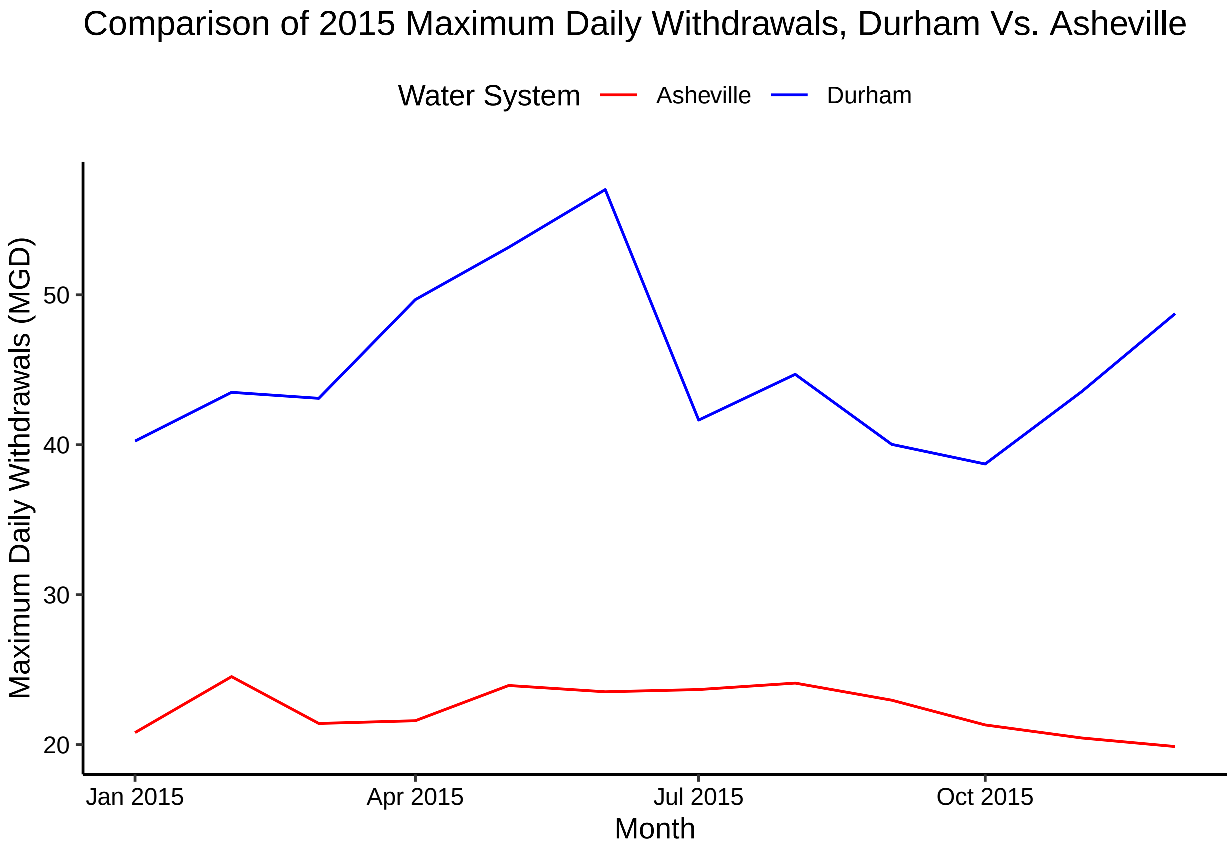


8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares Asheville's to Durham's water withdrawals.

```
#8
df.ash <- fun(2015, '01-11-010')

df_combined <- rbind(df_2015, df.ash)
df_combined
```

```
##    water.system.name     PWSID Year    ownership Month max.day.use       Date
## 1             Durham 03-32-010 2015 Municipality   Apr       49.68 2015-04-01
## 2             Durham 03-32-010 2015 Municipality   Aug       44.70 2015-08-01
## 3             Durham 03-32-010 2015 Municipality   Dec       48.75 2015-12-01
## 4             Durham 03-32-010 2015 Municipality   Feb       43.50 2015-02-01
## 5             Durham 03-32-010 2015 Municipality   Jan       40.25 2015-01-01
## 6             Durham 03-32-010 2015 Municipality   Jul       41.65 2015-07-01
## 7             Durham 03-32-010 2015 Municipality   Jun       57.02 2015-06-01
## 8             Durham 03-32-010 2015 Municipality   Mar       43.10 2015-03-01
## 9             Durham 03-32-010 2015 Municipality   May       53.17 2015-05-01
## 10            Durham 03-32-010 2015 Municipality   Nov       43.55 2015-11-01
## 11            Durham 03-32-010 2015 Municipality   Oct       38.72 2015-10-01
## 12            Durham 03-32-010 2015 Municipality   Sep       40.03 2015-09-01
## 13         Asheville 01-11-010 2015 Municipality   Apr       21.60 2015-04-01
## 14         Asheville 01-11-010 2015 Municipality   Aug       24.11 2015-08-01
## 15         Asheville 01-11-010 2015 Municipality   Dec       19.88 2015-12-01
## 16         Asheville 01-11-010 2015 Municipality   Feb       24.54 2015-02-01
## 17         Asheville 01-11-010 2015 Municipality   Jan       20.81 2015-01-01
## 18         Asheville 01-11-010 2015 Municipality   Jul       23.68 2015-07-01
## 19         Asheville 01-11-010 2015 Municipality   Jun       23.53 2015-06-01
## 20         Asheville 01-11-010 2015 Municipality   Mar       21.42 2015-03-01
## 21         Asheville 01-11-010 2015 Municipality   May       23.95 2015-05-01
## 22         Asheville 01-11-010 2015 Municipality   Nov       20.45 2015-11-01
## 23         Asheville 01-11-010 2015 Municipality   Oct       21.32 2015-10-01
## 24         Asheville 01-11-010 2015 Municipality   Sep       22.97 2015-09-01
```

```r
# Create plot
ggplot(df_combined, aes(x = Date, y = max.day.use, color = water.system.name)) +
  geom_line() +
  labs(title = "Comparison of 2015 Maximum Daily Withdrawals, Durham Vs. Asheville",
       x = "Month",
       y = "Maximum Daily Withdrawals (MGD)",
       color = "Water System") +
  scale_color_manual(values = c("red", "blue"))
```



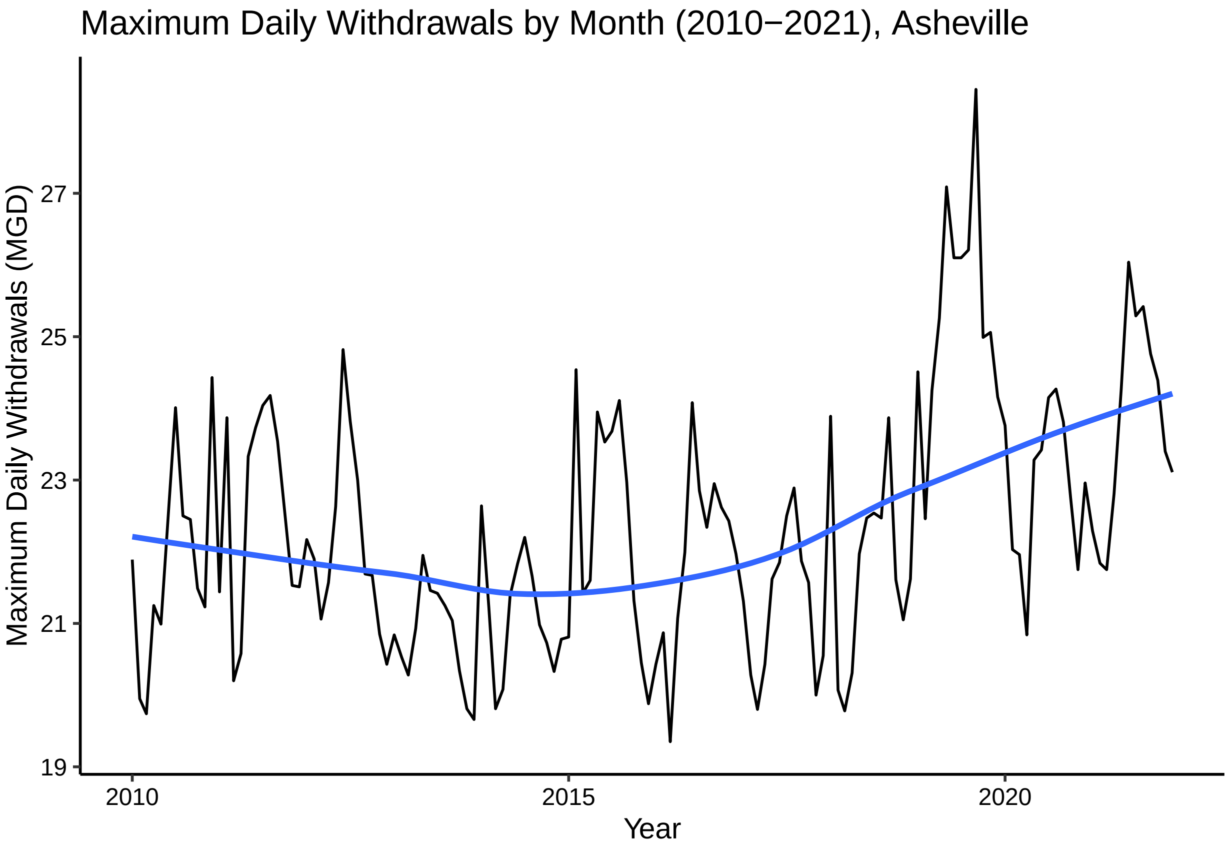
9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2021.Add a smoothed line to the plot (method = 'loess').

   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.

```r
#9
years <- 2010:2021
PWSIDs <- rep('01-11-010',length(years))
```

```
years.ash <- map2(years, PWSIDs, fun)

df_ashev <- bind_rows(years.ash)
df_ashev <- arrange(df_ashev, Date)

ggplot(df_ashev, aes(x = Date, y = max.day.use)) +
  geom_line() +
  geom_smooth(method = "loess", se = FALSE) +
  labs(title = "Maximum Daily Withdrawals by Month (2010-2021), Asheville",
       x = "Year",
       y = "Maximum Daily Withdrawals (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?

Yes, the maximum water usage trend shows a decrease from 2010 to 2015, followed by an increase from 2015 to 2020, with the usage ultimately surpassing the initial levels of 2010.