Assignment 09: Data Scraping

Katie Owens

Total points:

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on data scraping.

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Fay_09_Data_Scraping.Rmd") prior to submission.

Set up

1. Set up your session:

- Check your working directory
- Load the packages `tidyverse`, `rvest`, and any others you end up using.
- Set your ggplot theme

```
#1
setwd("C:/Users/Katherine/Documents/872-Data Analytics/Environmental_Data_Analytics_2022")
getwd()
```

```
## [1] "C:/Users/Katherine/Documents/872-Data Analytics/Environmental_Data_Analytics_2022"
```

```
library(tidyverse)
library(rvest)
```

```
## Warning: package 'rvest' was built under R version 4.1.3
```

```
library(lubridate)

#set ggplot theme
mytheme <- theme_classic(base_size = 12) +
  theme(axis.text = element_text(color = "purple"),
        legend.position = "bottom")
theme_set(mytheme)
```

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham's 2019 Municipal Local Water Supply Plan (LWSP):

- Navigate to https://www.ncwater.org/WUDC/app/LWSP/search.php
- Change the date from 2020 to 2019 in the upper right corner.

- 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=2020

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

```
#https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2020
#2
#Fetch the web resources from the URL
webpage1 <- read_html('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2020')
```

3. The data we want to collect are listed below:

- From the "1. System Information" section:
- Water system name
- PSWID
- Ownership
- From the "3. Water Supply Sources" section:
- MAX Daily Use (MGD) - for each month

In the code chunk below scrape these values, assigning them to three 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, with the first value being 36.0100.

```
#3
water.system.name <- webpage1 %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>% html_text()
water.system.name
```

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

```
pswid <- webpage1 %>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)") %>% html_text()
pswid
```

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

```
ownership <- webpage1 %>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>% html_text()
ownership
```

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

```
max.withdrawals.mgd <- webpage1 %>%
  html_nodes("th~ td+ td") %>%  html_text()
max.withdrawals.mgd
```

```
##  [1] "36.0100" "36.9800" "41.6900" "32.0500" "40.6100" "40.5600" "37.2900"
##  [8] "43.6300" "33.3200" "32.3700" "41.9300" "28.0600"
```

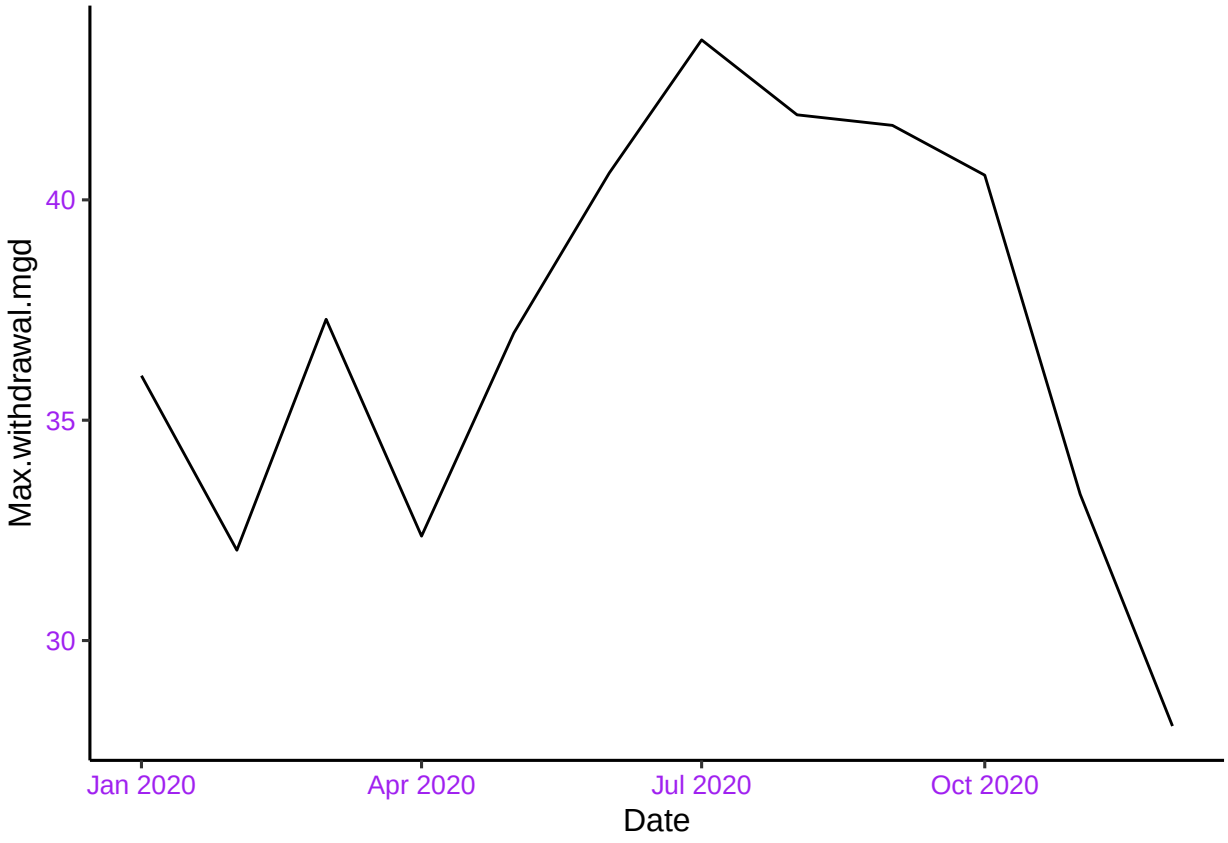
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   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 order. You can overcome this by creating a month column in the same order the data are scraped: Jan, May, Sept, Feb, etc...

5. Plot the max daily withdrawals across the months for 2020

```
#4
df1 <- data.frame(
  "Water.system.name" = rep(water.system.name,12), #make column with 12 rows and 1 column for water sys
  "PSWID" = rep(pswid,12), #make PSWIP column with 12 rows
  "Ownership" = rep(ownership,12), #results in 36 rows, bad
  "Max.withdrawal.mgd" = as.numeric(max.withdrawals.mgd),
  "Month" = c(1,5,9,2,6,10,3,7,11,4,8,12), #Telling R
  "Year" = rep(2020,12)) %>% #error with year
  mutate(Date = my(paste(Month,"-",Year)))


####parentheses highlight the  pair, underlines the one missing partner (no highlight)

#5
ggplot(df1, aes(x=Date, y=Max.withdrawal.mgd)) +
  geom_line()
```



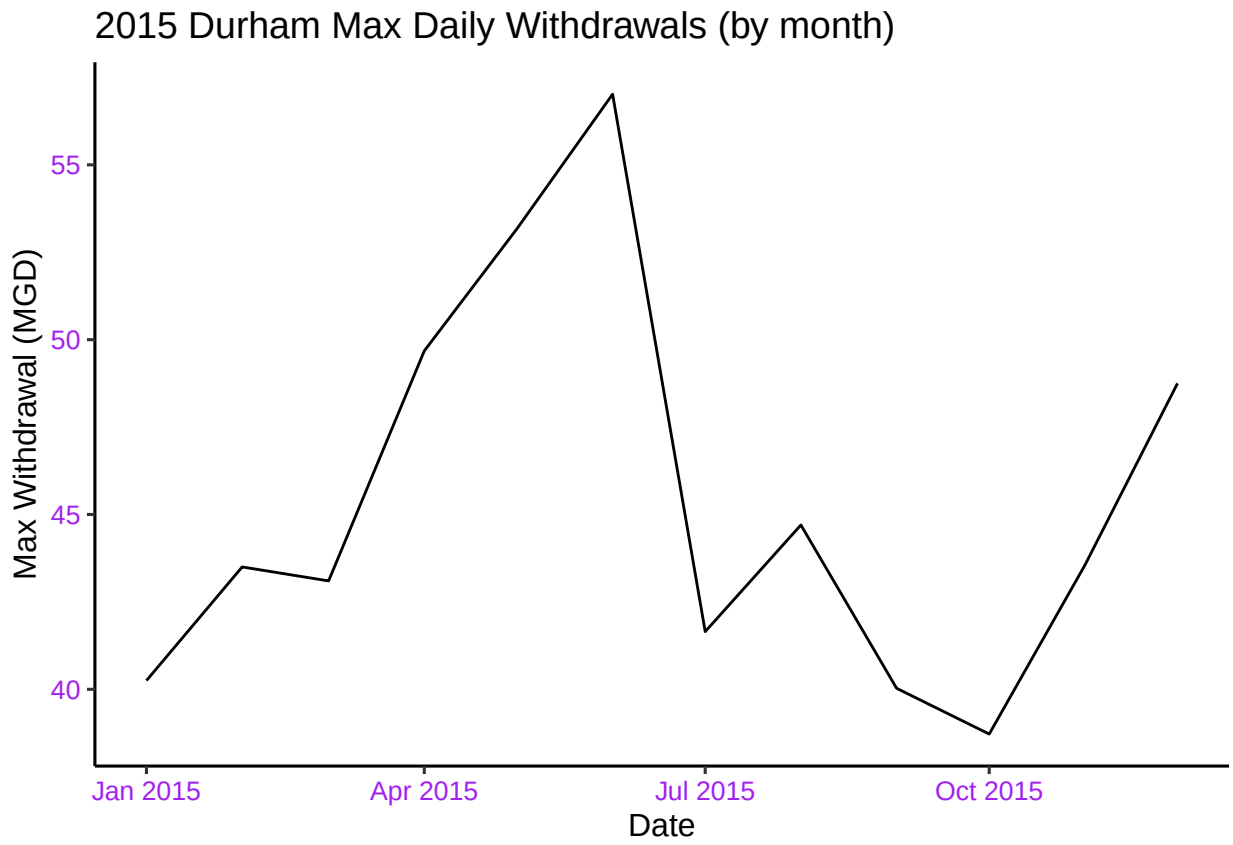
6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site scraped**.

```r
#6.
#Make function
NCW_scrape_fcn <- function(the_pwsid,the_year){

  #create website variable
  NCW_website <- read_html(paste0("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=",the_pwsid,

  #create element variables - correct
 water.system.name <- NCW_website %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>% html_text()
water.system.name

pswid <- NCW_website %>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)") %>% html_text()
pswid

ownership <- NCW_website %>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>% html_text()
ownership

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



  #scrape the data items
water.system.name <- NCW_website %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>% html_text()
water.system.name

pswid <- NCW_website %>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)") %>% html_text()
pswid

ownership <- NCW_website %>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>% html_text()
ownership

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

  #make scraped data frame
df1 <- data.frame(
  "Water.system.name" = rep(water.system.name,12), #make column with 12 rows and 1 column for water sys
  "PSWID" = rep(the_pwsid,12), #make PSWIP column with 12 rows with
  "Ownership" = rep(ownership,12),
  "Max.withdrawal.mgd" = as.numeric(max.withdrawals.mgd),
  "Month" = c(1,5,9,2,6,10,3,7,11,4,8,12), #Telling R which month is which
  "Year" = rep(the_year,12)) %>%
  mutate(Date = my(paste(Month,"-",Year)))
```

```
  #return the data frama
  return(df1)
}
```

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

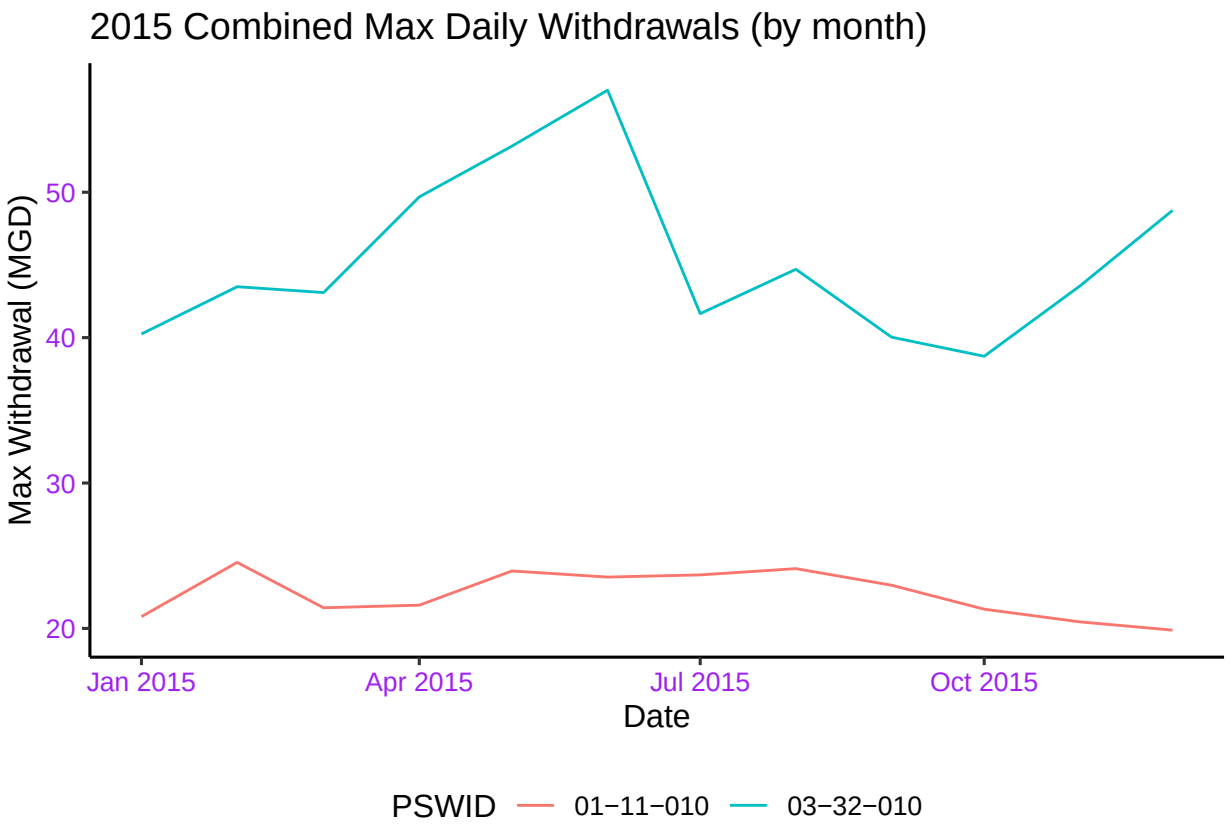
```
#7
Durham_2015 <- NCW_scrape_fcn("03-32-010", 2015)
view (Durham_2015)


ggplot(Durham_2015, aes(x=Date, y=Max.withdrawal.mgd)) +
  geom_line() +
  labs(x = "Date", y = "Max Withdrawal (MGD)",
       title = "2015 Durham Max Daily Withdrawals (by month)")
```



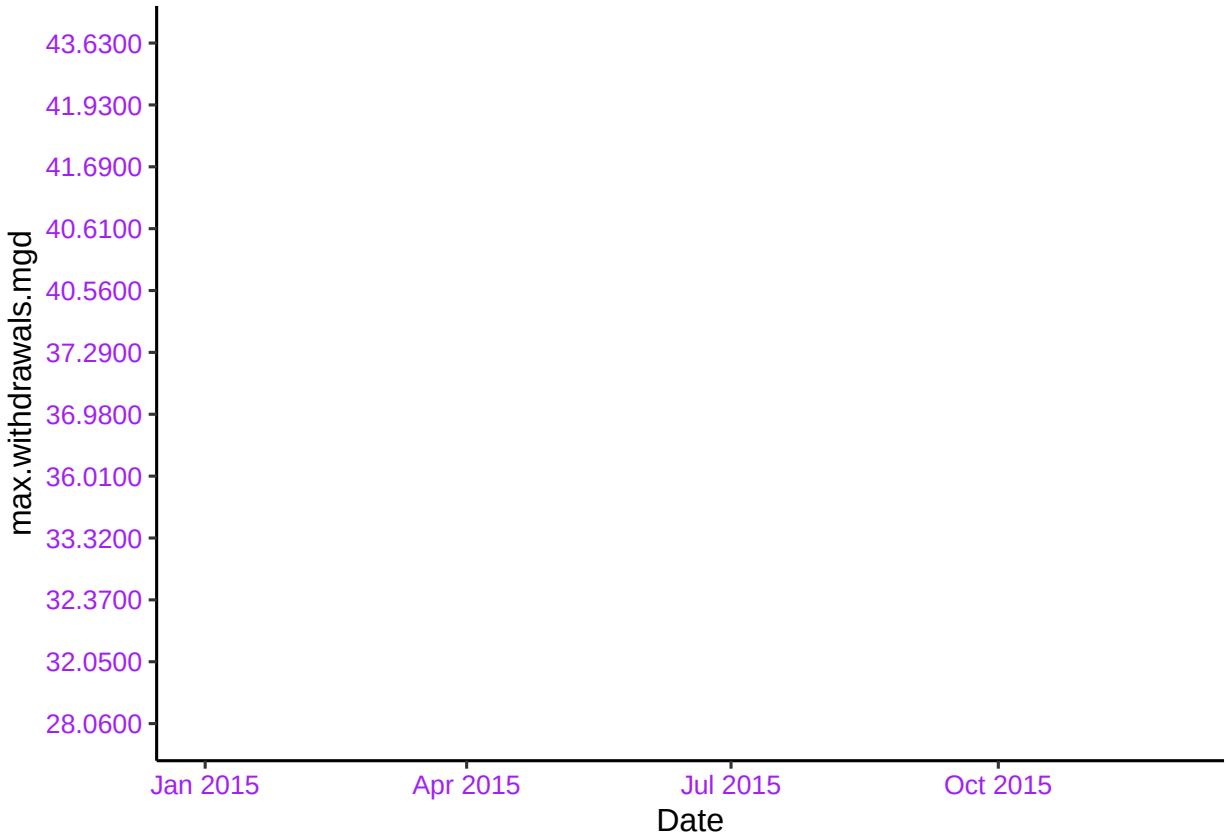
8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares the Asheville to Durham’s water withdrawals.

```
#8
Asheville_2015 <- NCW_scrape_fcn("01-11-010", 2015)
df_combined <- bind_rows(Durham_2015, Asheville_2015)

ggplot(df_combined, aes(color=PSWID, x=Date, y=Max.withdrawal.mgd)) +
```

```
  geom_line() +
  labs(x = "Date", y = "Max Withdrawal (MGD)",
       title = "2015 Combined Max Daily Withdrawals (by month)")
```



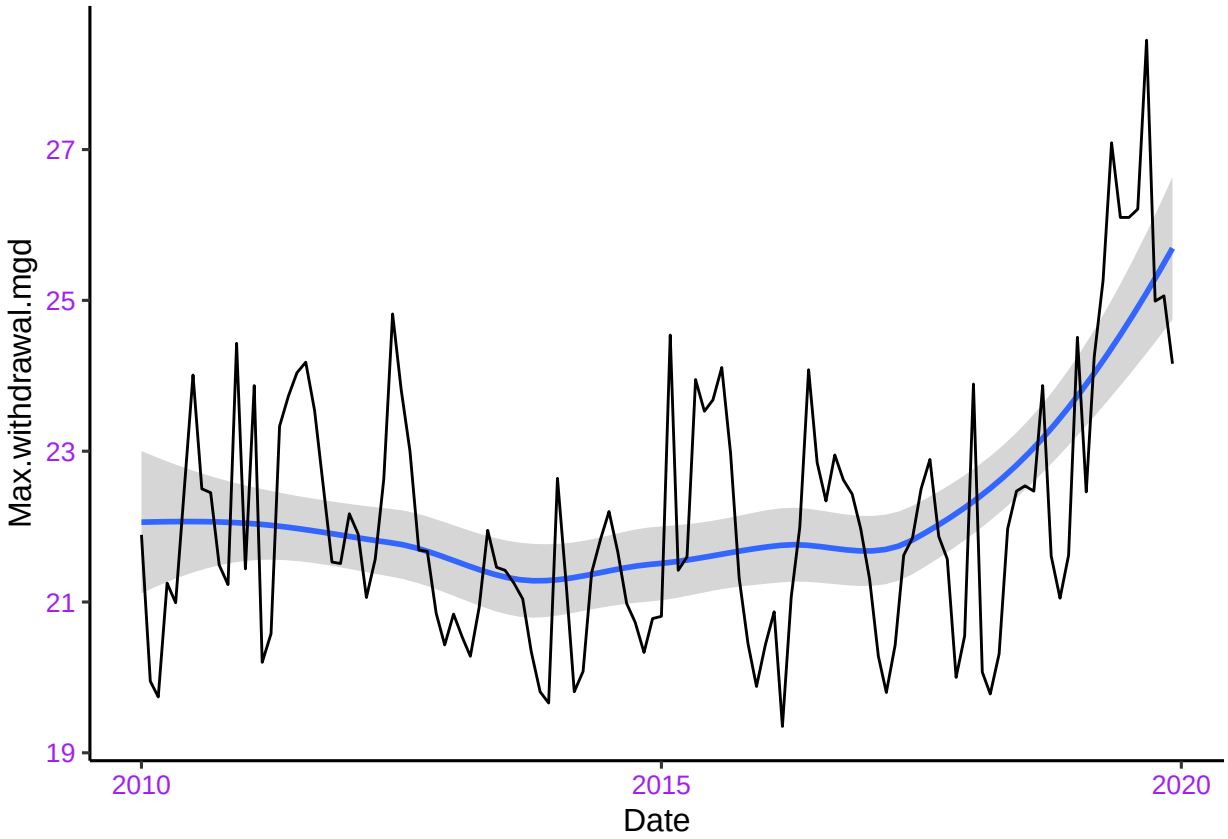

```
ggplot(data = subset(df_combined, PSWID == "03-32-010"), aes(x = Date, y = max.withdrawals.mgd))
```



9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

```
#9

#Make years set
the_years = seq(2010, 2019)
my_facility = '01-11-010'

Asheville2010to2019 <- the_years %>%
  map(NCW_scrape_fcn, the_pwsid = my_facility) %>%
  bind_rows()

ggplot(Asheville2010to2019,aes(x=Date,y=Max.withdrawal.mgd)) +
 geom_smooth(method="loess") +
 geom_line()
```

```
## `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, there is an increasing trend over time with average usage increasing by roughly 10% between 2017-2020.