Assignment 09: Data Scraping

Elise Boos

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
#check wd
getwd()
```

```
## [1] "/Users/elise/Desktop/Data_Analytics/Environmental_Data_Analytics_2022/Assignments"
```

```
#load packages
require(tidyverse)
require(rvest)
require(lubridate)
```

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.)

```
#2
#set webpage variable reading in as url
webpage <- read_html('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2020')
webpage
```

```
## {html_document}
## <html xmlns="http://www.w3.org/1999/xhtml" lang="en" xml:lang="en">
## [1] <head>\n<title>DWR :: Local Water Supply Planning</title>\n<meta http-equ ...
## [2] <body id="plan">\r\n<!--<div id="division-header">\r\n<a name="top" href= ...
```

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:

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

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

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

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

```
ownership <- webpage %>%
  html_nodes("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] "36.0100" "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

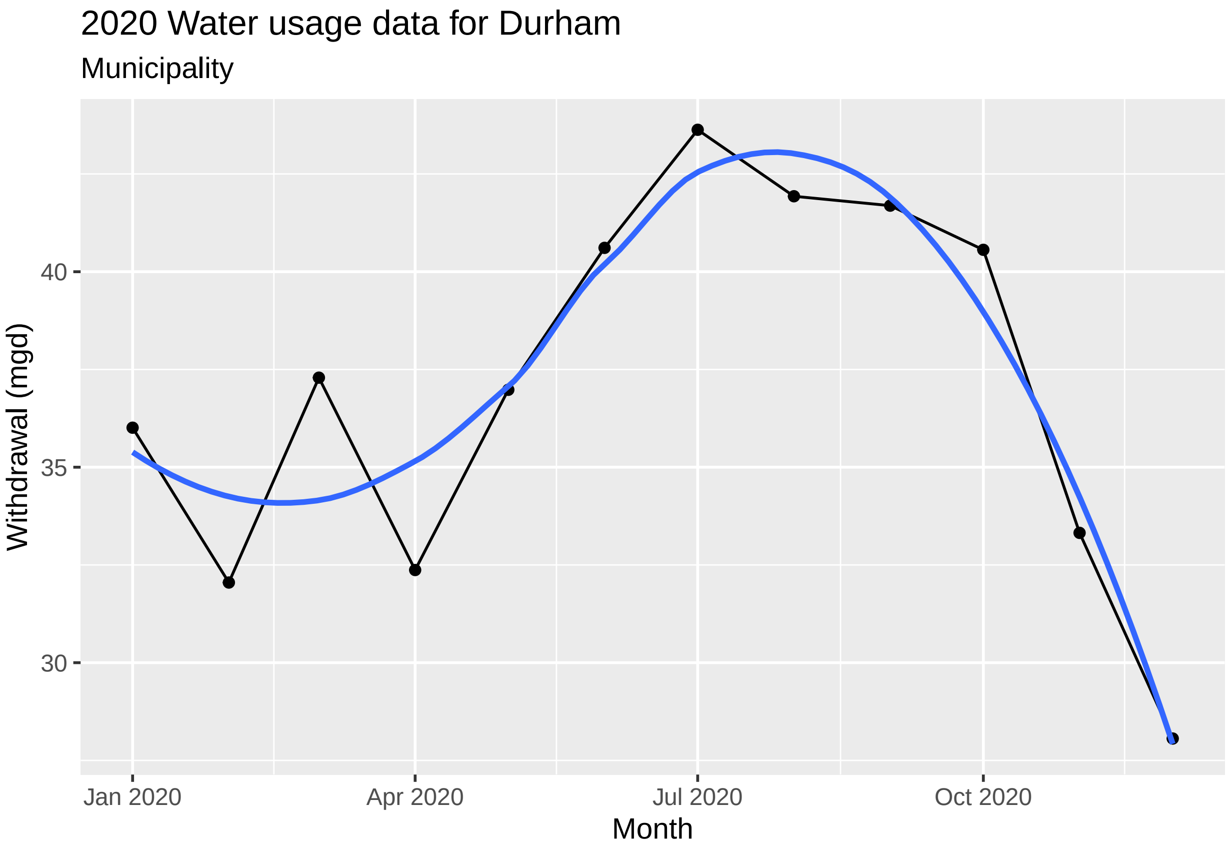
```
#4
#create dataframe from scaped data
withdrawals <- data.frame("Month" = month(c(1,5,9,2,6,10,3,7,11,4,8,12)),
                          "Year" = rep(2020,12),
                          "Water.System" = water.system.name,
                          "PSWID" = pswid,
                          "Ownership" = ownership,
                          "Max.Withdrawals.mgd" = as.numeric(max.withdrawals.mgd))
withdrawals_2020 <- withdrawals %>%
  mutate(Date = my(paste0(Month,"-", Year)))

withdrawals_2020
```

```
##    Month Year Water.System     PSWID    Ownership Max.Withdrawals.mgd
## 1      1 2020       Durham 03-32-010 Municipality               36.01
## 2      5 2020       Durham 03-32-010 Municipality               36.98
## 3      9 2020       Durham 03-32-010 Municipality               41.69
## 4      2 2020       Durham 03-32-010 Municipality               32.05
## 5      6 2020       Durham 03-32-010 Municipality               40.61
## 6     10 2020       Durham 03-32-010 Municipality               40.56
## 7      3 2020       Durham 03-32-010 Municipality               37.29
## 8      7 2020       Durham 03-32-010 Municipality               43.63
## 9     11 2020       Durham 03-32-010 Municipality               33.32
## 10     4 2020       Durham 03-32-010 Municipality               32.37
## 11     8 2020       Durham 03-32-010 Municipality               41.93
## 12    12 2020       Durham 03-32-010 Municipality               28.06
##          Date
## 1  2020-01-01
## 2  2020-05-01
## 3  2020-09-01
## 4  2020-02-01
## 5  2020-06-01
## 6  2020-10-01
## 7  2020-03-01
## 8  2020-07-01
## 9  2020-11-01
## 10 2020-04-01
## 11 2020-08-01
## 12 2020-12-01
```

```
#5
#plot max withdrawals by month
ggplot(withdrawals_2020,aes(x=Date,y=Max.Withdrawals.mgd)) +
  geom_point() +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2020 Water usage data for",water.system.name),
       subtitle = ownership,
       y="Withdrawal (mgd)",
       x="Month")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



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**.

```
#6.
#create scrape it function
scrape.it <- function(the_year, pswid_year){
  #Get the proper url
  the_url <- read_html(
    paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=', pswid_year,'&year=',the_year))

  #Fetch the website
water.system.name <- the_url %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()
pswid <- the_url %>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
  html_text()
ownership <- the_url %>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%
  html_text()
max.withdrawals.mgd <- the_url %>%
  html_nodes("th~ td+ td") %>%
  html_text()

  #Construct a dataframe from the values
  the_df <- tibble("Month" = month(c(1,5,9,2,6,10,3,7,11,4,8,12)),
                     "Water.System" = water.system.name,
                     "PSWID" = pswid,
                     "Ownership" = ownership,
                     "Max.Withdrawals.mgd" = as.numeric(max.withdrawals.mgd)) %>%
   mutate(Year = the_year) %>%
   mutate(Date = my(paste0(Month,"-", Year)))

  return(the_df)
}
```

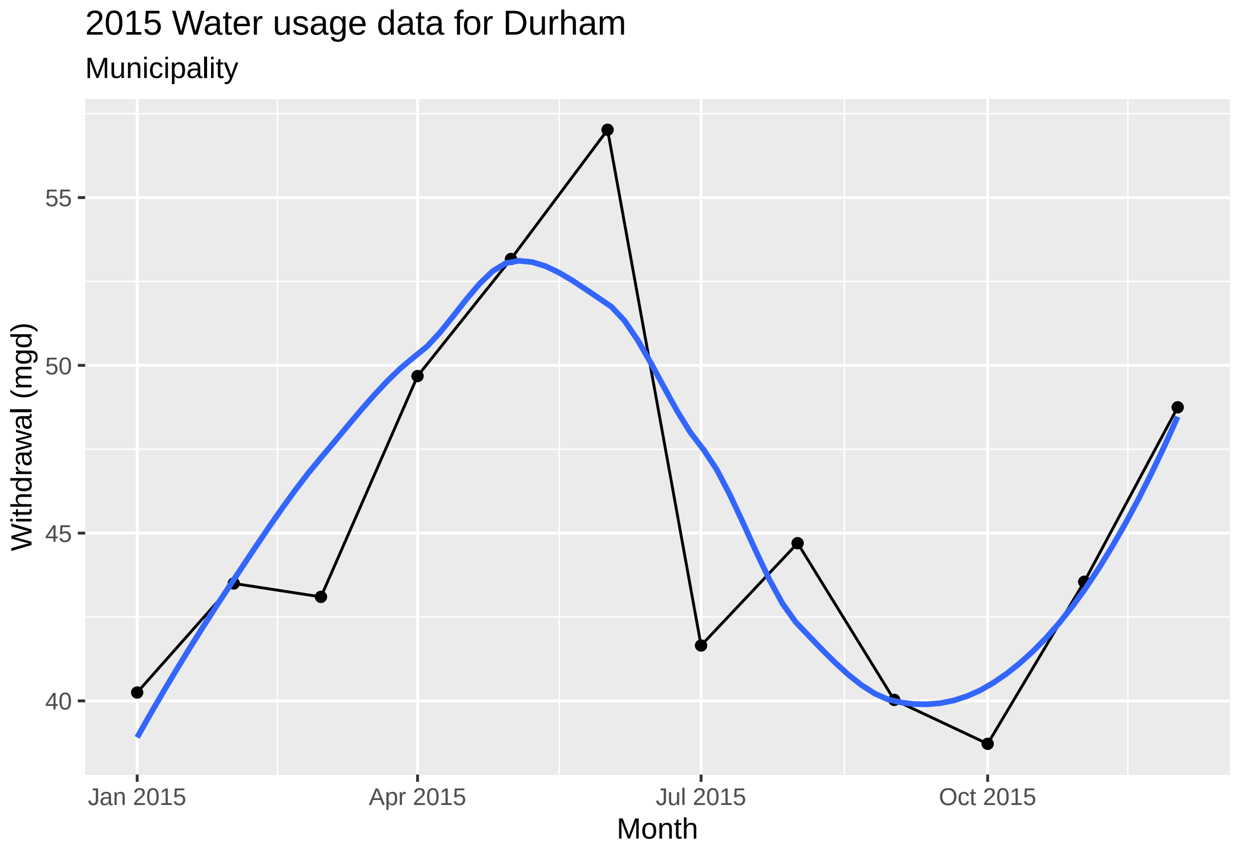
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID=‘03-32-010’) for each month in 2015

```
#7
#use function for durham and 2015
durham_2015 <- scrape.it(2015, '03-32-010')
durham_2015
```

```
## # A tibble: 12 x 7
##    Month Water.System PSWID     Ownership    Max.Withdrawals.mgd  Year Date
##    <dbl> <chr>        <chr>     <chr>                      <dbl> <dbl> <date>
##  1     1 Durham       03-32-010 Municipality                40.2  2015 2015-01-01
##  2     5 Durham       03-32-010 Municipality                53.2  2015 2015-05-01
##  3     9 Durham       03-32-010 Municipality                40.0  2015 2015-09-01
##  4     2 Durham       03-32-010 Municipality                43.5  2015 2015-02-01
##  5     6 Durham       03-32-010 Municipality                57.0  2015 2015-06-01
##  6    10 Durham       03-32-010 Municipality                38.7  2015 2015-10-01
##  7     3 Durham       03-32-010 Municipality                43.1  2015 2015-03-01
##  8     7 Durham       03-32-010 Municipality                41.6  2015 2015-07-01
##  9    11 Durham       03-32-010 Municipality                43.6  2015 2015-11-01
## 10     4 Durham       03-32-010 Municipality                49.7  2015 2015-04-01
## 11     8 Durham       03-32-010 Municipality                44.7  2015 2015-08-01
## 12    12 Durham       03-32-010 Municipality                48.8  2015 2015-12-01
```

```
#plot max withdrawals by month
ggplot(durham_2015,aes(x=Date,y=Max.Withdrawals.mgd)) +
  geom_point() +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2015 Water usage data for Durham"),
       subtitle = ownership,
       y="Withdrawal (mgd)",
       x="Month")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



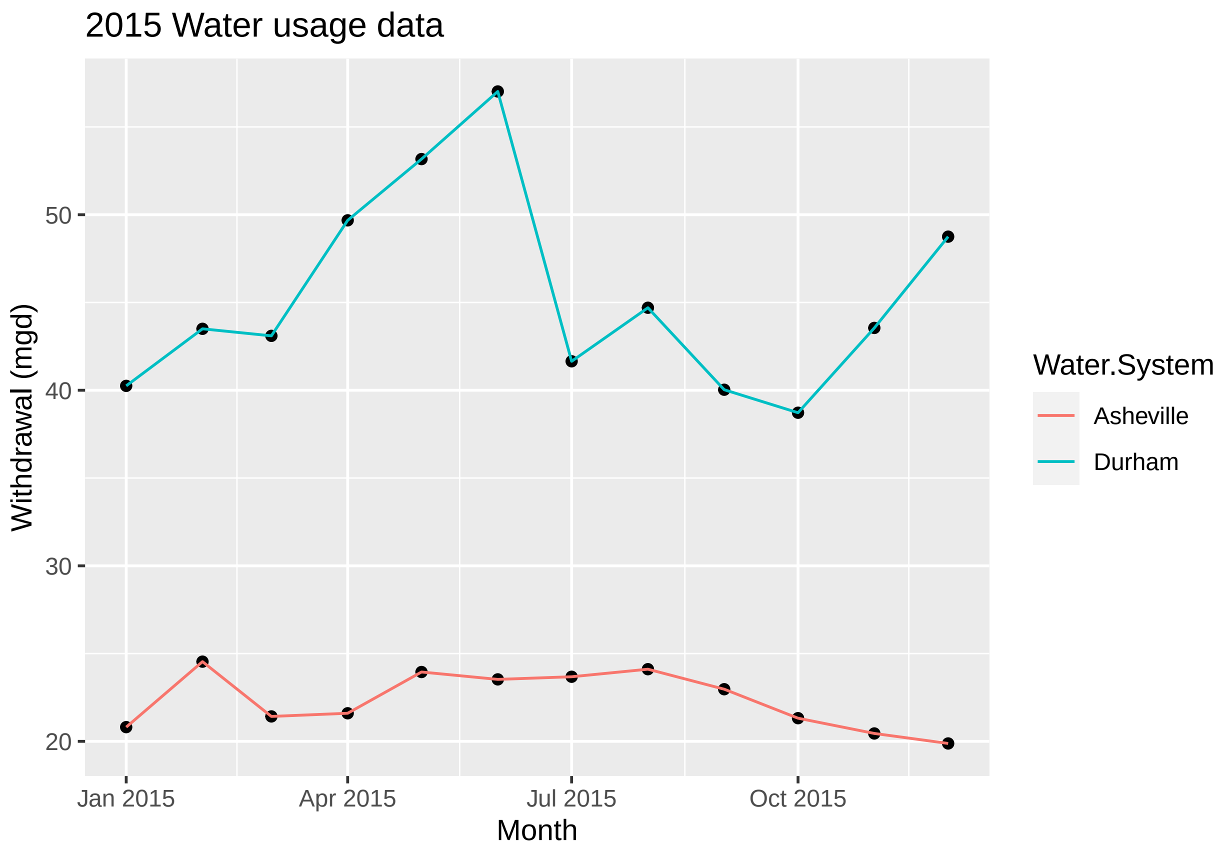
8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares the Asheville to Durham's water withdrawals.

```
#8
#asheville for 2015
asheville_2015 <- scrape.it(2015, '01-11-010')
asheville_2015
```

```
## # A tibble: 12 x 7
##    Month Water.System PSWID     Ownership    Max.Withdrawals.mgd  Year Date      
##    <dbl> <chr>        <chr>     <chr>                      <dbl> <dbl> <date>    
##  1     1 Asheville    01-11-010 Municipality                20.8  2015 2015-01-01
##  2     5 Asheville    01-11-010 Municipality                24.0  2015 2015-05-01
##  3     9 Asheville    01-11-010 Municipality                23.0  2015 2015-09-01
##  4     2 Asheville    01-11-010 Municipality                24.5  2015 2015-02-01
##  5     6 Asheville    01-11-010 Municipality                23.5  2015 2015-06-01
##  6    10 Asheville    01-11-010 Municipality                21.3  2015 2015-10-01
##  7     3 Asheville    01-11-010 Municipality                21.4  2015 2015-03-01
##  8     7 Asheville    01-11-010 Municipality                23.7  2015 2015-07-01
##  9    11 Asheville    01-11-010 Municipality                20.4  2015 2015-11-01
## 10     4 Asheville    01-11-010 Municipality                21.6  2015 2015-04-01
## 11     8 Asheville    01-11-010 Municipality                24.1  2015 2015-08-01
## 12    12 Asheville    01-11-010 Municipality                19.9  2015 2015-12-01
```

```
#bind durham and asheville dataframes
a_d_2015 <- rbind(durham_2015, asheville_2015)

#plot coloring by water system
ggplot(a_d_2015,aes(x=Date,y=Max.Withdrawals.mgd)) +
  geom_point() +
  geom_line(aes(color = Water.System)) +
  labs(title = paste("2015 Water usage data"),
       y="Withdrawal (mgd)",
       x="Month")
```

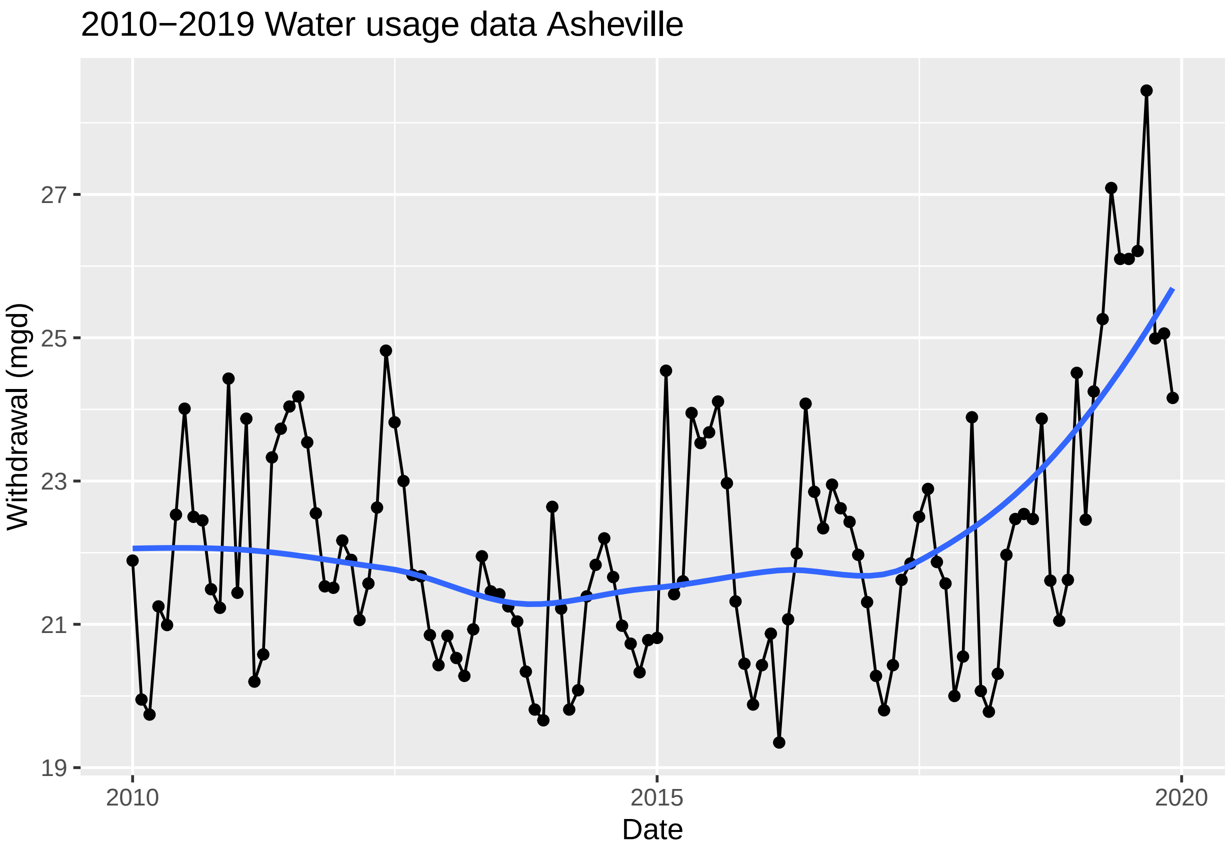


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

```
#9
#use lapply to run for 2010-2019
years <- seq(2010,2019)
asheville_2010_2019 <- lapply(X = years, FUN = scrape.it, pswid_year = '01-11-010') %>% bind_rows()

#plot
ggplot(asheville_2010_2019,aes(x=Date,y=Max.Withdrawals.mgd)) +
  geom_point() +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2010-2019 Water usage data Asheville"),
       y="Withdrawal (mgd)",
       x="Date")
```

```
## `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 over time there is an increase in water usage in Asheville.