Assignment 10: Data Scraping

Jess Garcia

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_10_Data_Scraping.Rmd") prior to submission.

The completed exercise is due on Tuesday, April 6 at 11:59 pm.

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 working directory, load packages, & set theme
getwd()
```

```
## [1] "C:/Users/93jes/Documents/ENV872/Environmental_Data_Analytics_2021"
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.4
```

```
library(rvest)
library(lubridate)



JG_theme <-
theme_bw(base_size = 12)+
theme(axis.text = element_text(color = "black"),
plot.title = element_text(color = "black", size = 16, face = "bold",
hjust = 0.5))
```

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=2019

Indicate this website as the as the URL to be scraped.

```
#2. Website scraping URL

Durham_2019.webpage <- read_html ('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&yea
```

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

- From the "System Information" section:

- Water system name

- PSWID

- Ownership

- From the "Water Supply Sources" section:

- Maximum monthly withdrawals (MGD)

In the code chunk below scrape these values into the supplied variable names.

```
#3. Create variables with scraped data

Water_System_Name <-
  Durham_2019.webpage%>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()
Water_System_Name
```

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

```
PSWID <-
  Durham_2019.webpage%>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
  html_text()
PSWID
```

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

```
Ownership <-
  Durham_2019.webpage%>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%
  html_text()
Ownership
```

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

```
Max_monthly_withdraw<-
  Durham_2019.webpage%>%
  html_nodes("th~ td+ td") %>%
  html_text()
Max_monthly_withdraw
```

```
##  [1] "29.6200" "35.7300" "54.0700" "32.3900" "37.8600" "44.3500" "36.4300"
##  [8] "46.0200" "36.0600" "32.6000" "42.0500" "31.2000"
```

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.)

   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 2019.

```
#4. Create dataframe

withdrawals.df <-
  data.frame("Water System Name" = Water_System_Name,
            "PSWID" = PSWID,
            "Ownership" = Ownership,
            "Month" = c(1,5,9,2,6,10,3,7,11,4,8,12),
            "Year" = rep(2019,12),
            "Max Monthly Withdraw.MGD" = as.numeric(Max_monthly_withdraw))

withdrawals.df
```

```
##    Water.System.Name     PSWID    Ownership Month Year Max.Monthly.Withdraw.MGD
## 1             Durham 03-32-010 Municipality     1 2019                    29.62
## 2             Durham 03-32-010 Municipality     5 2019                    35.73
## 3             Durham 03-32-010 Municipality     9 2019                    54.07
## 4             Durham 03-32-010 Municipality     2 2019                    32.39
## 5             Durham 03-32-010 Municipality     6 2019                    37.86
## 6             Durham 03-32-010 Municipality    10 2019                    44.35
## 7             Durham 03-32-010 Municipality     3 2019                    36.43
## 8             Durham 03-32-010 Municipality     7 2019                    46.02
## 9             Durham 03-32-010 Municipality    11 2019                    36.06
## 10            Durham 03-32-010 Municipality     4 2019                    32.60
## 11            Durham 03-32-010 Municipality     8 2019                    42.05
## 12            Durham 03-32-010 Municipality    12 2019                    31.20
```

```
#To reorder the dataframe by month after re-ordering the numbers to match the withdrawals
withdrawals2.df <-withdrawals.df[order(withdrawals.df$Month),]

withdrawals2.df
```

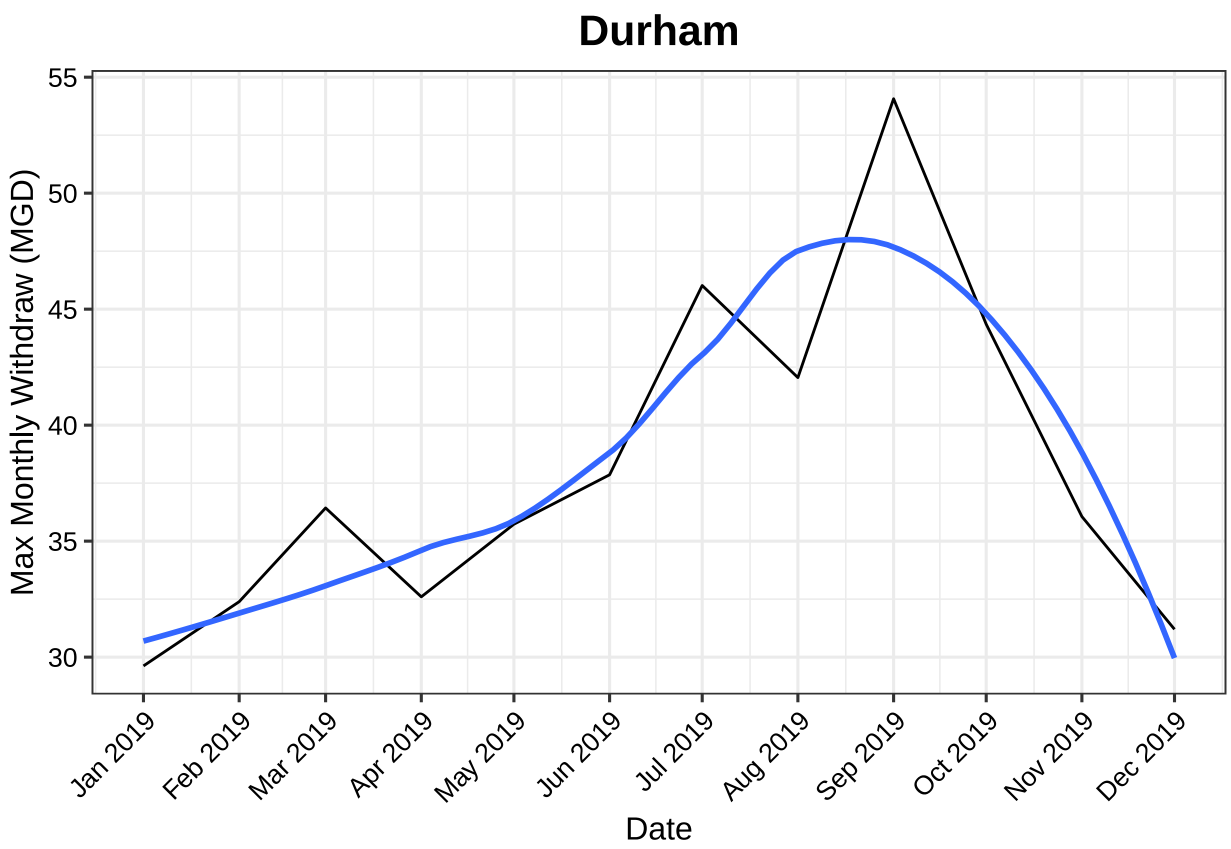
```
##    Water.System.Name     PSWID    Ownership Month Year Max.Monthly.Withdraw.MGD
## 1             Durham 03-32-010 Municipality     1 2019                    29.62
## 4             Durham 03-32-010 Municipality     2 2019                    32.39
## 7             Durham 03-32-010 Municipality     3 2019                    36.43
## 10            Durham 03-32-010 Municipality     4 2019                    32.60
## 2             Durham 03-32-010 Municipality     5 2019                    35.73
## 5             Durham 03-32-010 Municipality     6 2019                    37.86
## 8             Durham 03-32-010 Municipality     7 2019                    46.02
## 11            Durham 03-32-010 Municipality     8 2019                    42.05
## 3             Durham 03-32-010 Municipality     9 2019                    54.07
## 6             Durham 03-32-010 Municipality    10 2019                    44.35
## 9             Durham 03-32-010 Municipality    11 2019                    36.06
## 12            Durham 03-32-010 Municipality    12 2019                    31.20
```

```
withdrawals3.df <-
  withdrawals2.df %>%
  mutate(Date = (my(paste(Month, "-", Year))))

withdrawals3.df
```

```
##    Water.System.Name     PSWID    Ownership Month Year Max.Monthly.Withdraw.MGD
## 1             Durham 03-32-010 Municipality     1 2019                    29.62
## 4             Durham 03-32-010 Municipality     2 2019                    32.39
## 7             Durham 03-32-010 Municipality     3 2019                    36.43
## 10            Durham 03-32-010 Municipality     4 2019                    32.60
## 2             Durham 03-32-010 Municipality     5 2019                    35.73
## 5             Durham 03-32-010 Municipality     6 2019                    37.86
## 8             Durham 03-32-010 Municipality     7 2019                    46.02
## 11            Durham 03-32-010 Municipality     8 2019                    42.05
## 3             Durham 03-32-010 Municipality     9 2019                    54.07
## 6             Durham 03-32-010 Municipality    10 2019                    44.35
## 9             Durham 03-32-010 Municipality    11 2019                    36.06
## 12            Durham 03-32-010 Municipality    12 2019                    31.20
##          Date
## 1  2019-01-01
## 4  2019-02-01
## 7  2019-03-01
## 10 2019-04-01
## 2  2019-05-01
## 5  2019-06-01
## 8  2019-07-01
## 11 2019-08-01
## 3  2019-09-01
## 6  2019-10-01
## 9  2019-11-01
## 12 2019-12-01
```

```
#class(withdrawals3.df$Date)


#5. Plot max daily withdrawals across 2019

MaxWithdrawals_Durham_2019.plot <-
  ggplot(withdrawals3.df, aes(x = Date, y = Max.Monthly.Withdraw.MGD))+
  geom_line()+
  geom_smooth(method = "loess", se = FALSE)+
  ggtitle("Durham")+
  JG_theme+
  ylab("Max Monthly Withdraw (MGD)")+
  scale_x_date(date_breaks = "1 month", date_labels = "%b %Y")+
  theme(axis.text.x = element_text(angle=45, hjust = 1))

plot(MaxWithdrawals_Durham_2019.plot)
```

```
## `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 data scraped.

```
#6. Construct a data scraping function
base_url <-'https://www.ncwater.org/WUDC/app/LWSP/report.php?'

the_PWSID <- '03-32-010'

the_year <- '2019'

the_scrape_URL <- paste0(base_url,'pwsid=',the_PWSID,'&year=',the_year)

print(the_scrape_URL)
```

```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2019"
```

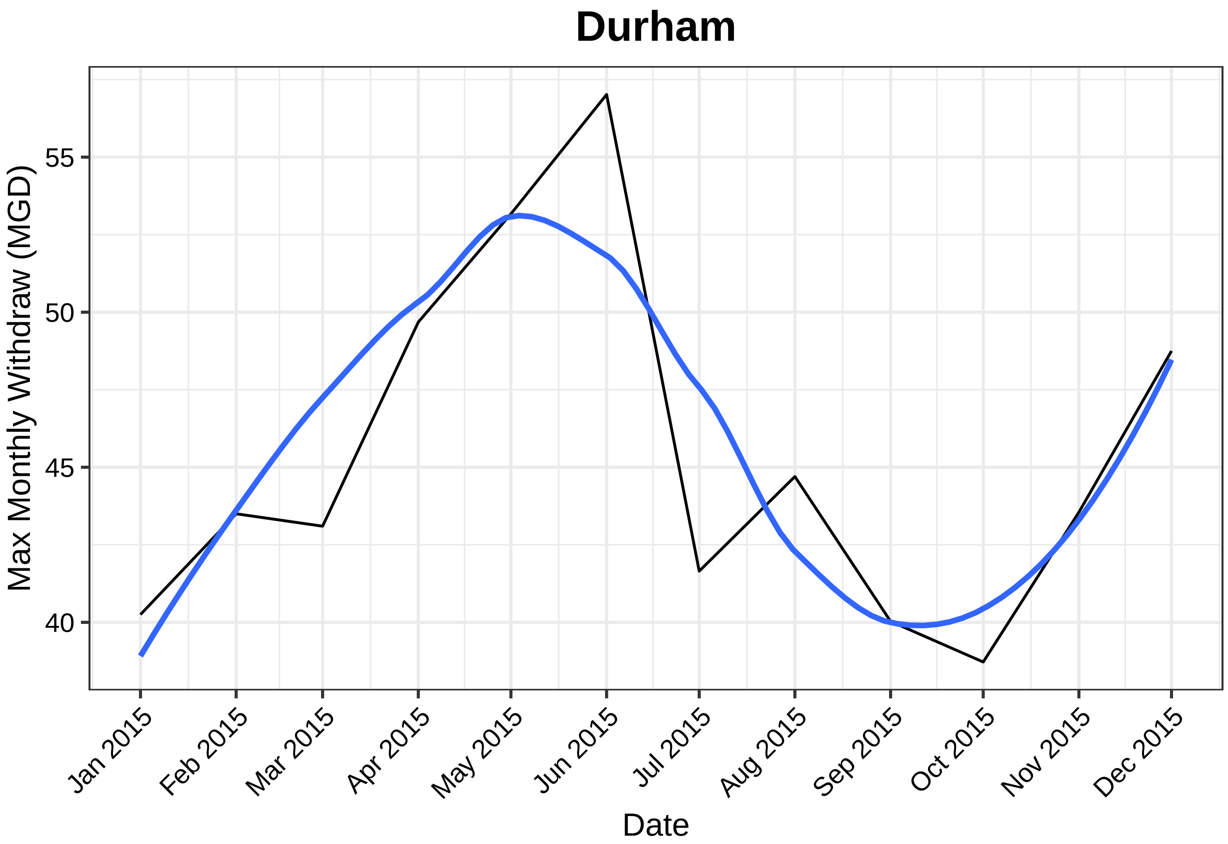
```
#Construct a function to scrape data for any PWSID & year
scrape.it <- function(the_year, the_PWSID){
  #Find the website
  base_url <-'https://www.ncwater.org/WUDC/app/LWSP/report.php?'

  the_scrape_URL <- paste0(base_url,'pwsid=',the_PWSID,'&year=',the_year)

  the_website <- read_html(the_scrape_URL)

  #Scrape the data
```

```
 Water_System_Name2 <-
  the_website%>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()
Water_System_Name2

PSWID2 <-
   the_website%>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
  html_text()
PSWID2

Ownership2 <-
  the_website%>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%
  html_text()
Ownership2

Max_monthly_withdraw2<-
  the_website%>%
  html_nodes("th~ td+ td") %>%
  html_text()
Max_monthly_withdraw2

#Convert to dataframe
withdrawals_function.df <-
  data.frame("Water System Name" = Water_System_Name2,
             "PSWID" = PSWID2,
             "Ownership" = Ownership2,
             "Month" = c(1,5,9,2,6,10,3,7,11,4,8,12),
             "Year" = the_year,
             "Max Monthly Withdraw.MGD" = as.numeric(Max_monthly_withdraw2))

withdrawals_function.df

#To reorder the dataframe by month after re-ordering the numbers to match the withdrawals
withdrawals_function2.df <-
  withdrawals_function.df[order(withdrawals.df$Month),]

withdrawals_function2.df


withdrawals_function3.df <-
  withdrawals_function2.df %>%
  mutate(Date = (my(paste(Month, "-", Year))))

withdrawals_function3.df

#Return the dataframe
return(withdrawals_function3.df)
}
```

7. Use the function above to extract and plot max daily withdrawals for Durham for each month in 2015

```
#7. Use the function to extract the data

#Durham_test <- scrape.it(2019, "03-32-010")

Durham_2015 <- scrape.it(2015,"03-32-010")

#Plot the Durham 2015 data
MaxWithdrawals_Durham_2015.plot <-
  ggplot(Durham_2015, aes(x = Date, y = Max.Monthly.Withdraw.MGD))+
  geom_line()+
  geom_smooth(method = "loess", se = FALSE)+
  ggtitle("Durham")+
  JG_theme+
  ylab("Max Monthly Withdraw (MGD)")+
  scale_x_date(date_breaks = "1 month", date_labels = "%b %Y")+
  theme(axis.text.x = element_text(angle=45, hjust = 1))

plot(MaxWithdrawals_Durham_2015.plot)
```

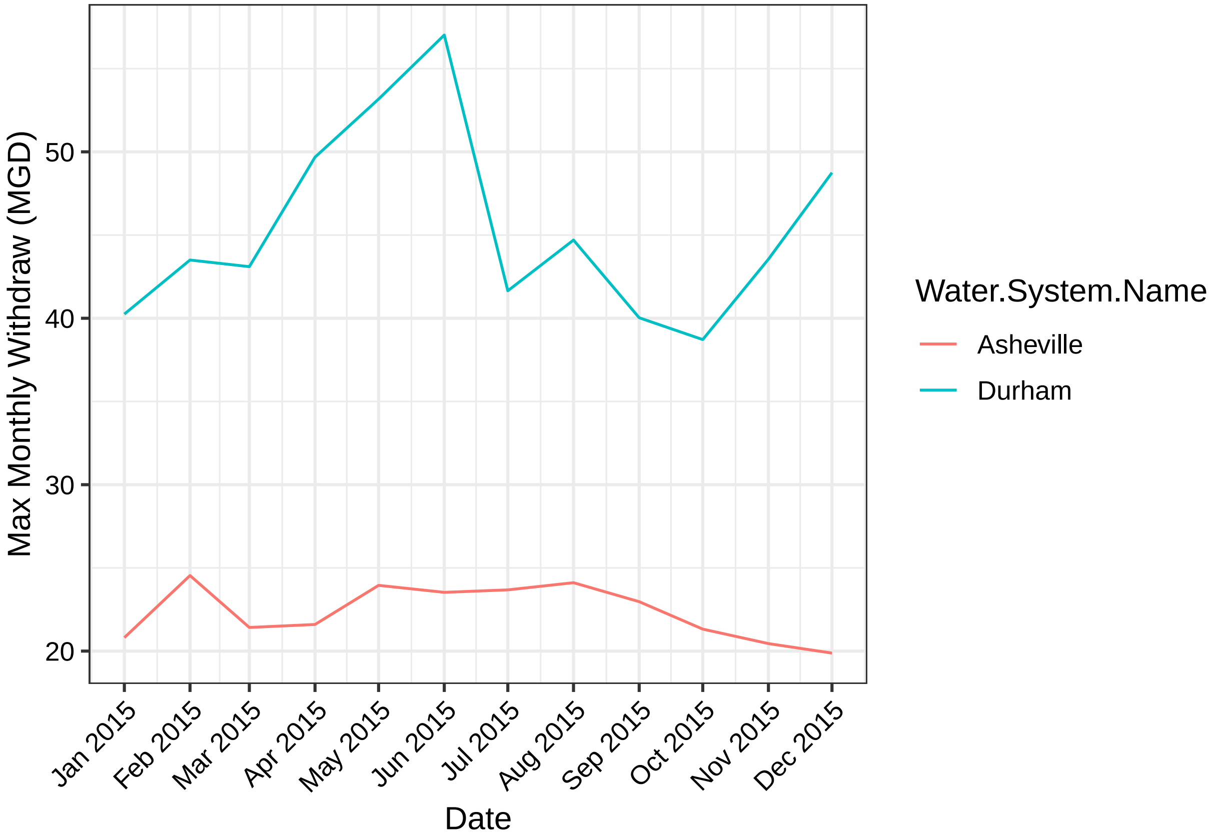
```
## `geom_smooth()` using formula 'y ~ x'
```



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. Use above function to scrape data for Asheville (PWSID = 01-11-010) in 2015
```

```
#Extract Ashevile data
Asheville_2015 <- scrape.it(2015, "01-11-010")

#Combine Durham & Asheville's withdrawals

Asheville_Durham_2015 <- rbind(Durham_2015, Asheville_2015)

#Plot Durham & Asheville withdrawals
Asheville_Durham_2015.plot <-
  ggplot(Asheville_Durham_2015, aes(x=Date, y=Max.Monthly.Withdraw.MGD, color=Water.System.Name))+
  geom_line()+
  JG_theme+
  ylab("Max Monthly Withdraw (MGD)")+
  scale_x_date(date_breaks = "1 month", date_labels = "%b %Y")+
  theme(axis.text.x = element_text(angle=45, hjust = 1))
Asheville_Durham_2015.plot
```

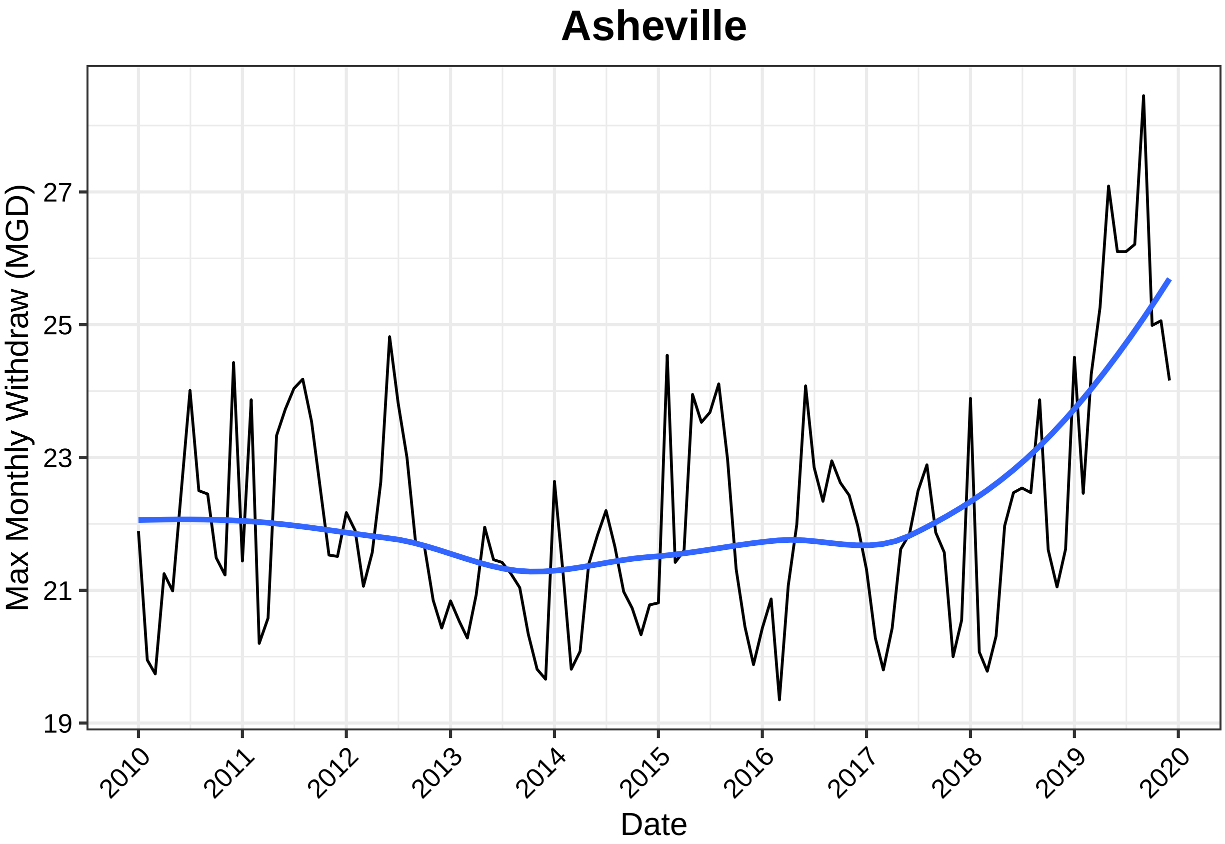


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

```
#9. Plot Asheville's max withdrawals by month for 2010-2019
Asheville_10yrwithdrawals <-
  map(rep(2010:2019),scrape.it,the_PWSID = "01-11-010") %>%
  bind_rows()
```

```
Asheville_10yrwithdrawals.plot <-
  ggplot(Asheville_10yrwithdrawals, aes(x= Date, y= Max.Monthly.Withdraw.MGD))+
  geom_line()+
  geom_smooth(method = "loess", se = FALSE)+
  JG_theme+
  ylab("Max Monthly Withdraw (MGD)")+
  scale_x_date(date_breaks = "1 year", date_labels = "%Y")+
  theme(axis.text.x = element_text(angle=45, hjust = 1))+
  ggtitle("Asheville")

Asheville_10yrwithdrawals.plot
```

```
## `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?

Just looking at the plot, it does look like Asheville has a trend of increased water withdrawals over time, particularly between 2017 and the present. Prior to 2017 the annual max withdrawals stayed somewhat the same.