Assignment 09: Data Scraping

Shidi Dai

OVERVIEW

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

Directions

1. Rename this file `<FirstLast>_A09_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 to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.

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
getwd()
```

```
## [1] "/home/guest/R/EDA-Fall2022"
```

```
library(tidyverse)
library(lubridate)

#install.packages("rvest")
library(rvest)

# 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 2021 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=2021

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=2021')
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:
- Maximum Daily 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_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()

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

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

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

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.)

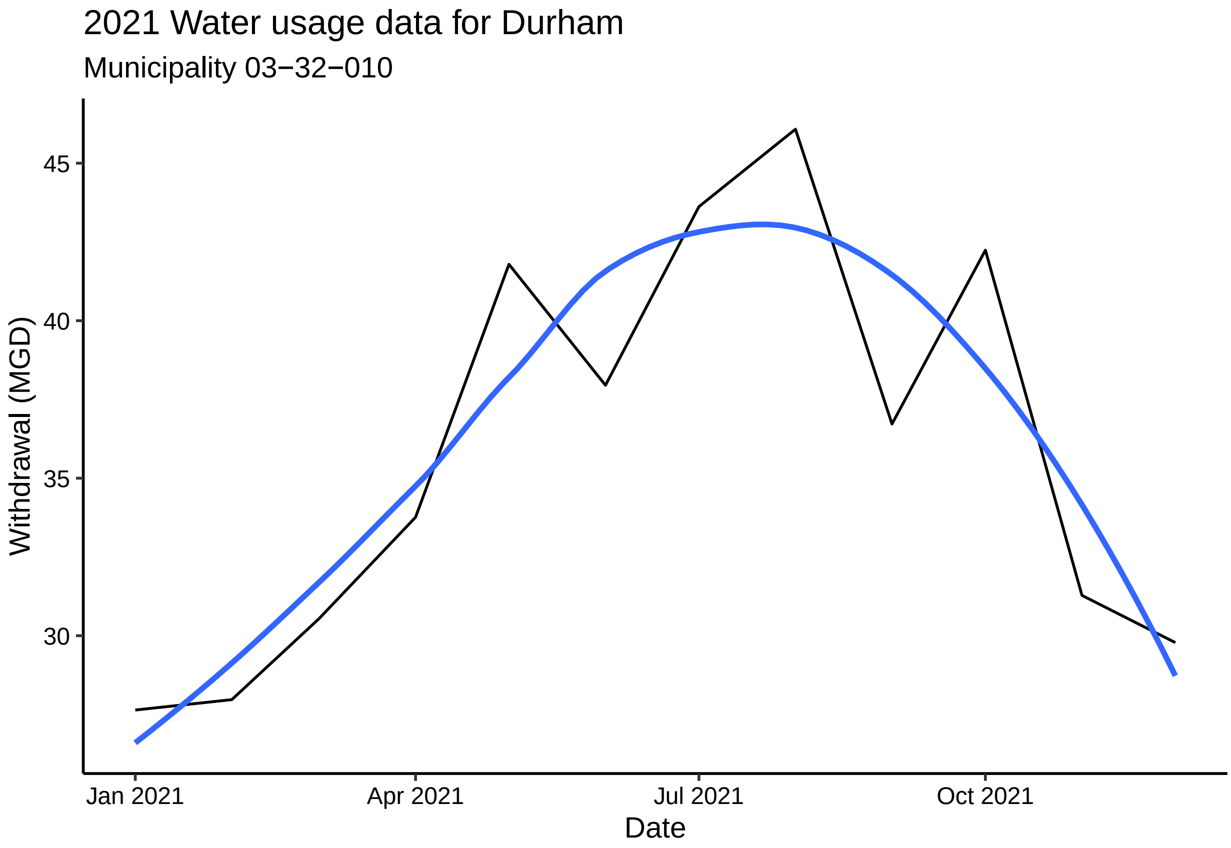
   TIP: Use `rep()` to repeat a value when creating a dataframe.

   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…

5. Create a line plot of the maximum daily withdrawals across the months for 2021

```
#4
df_withdrawals <- data.frame("Month" = c("01", "05", "09", "02", "06", "10",
                                          "03", "07", "11", "04", "08", "12"),
                              "Year" = rep(2021,12),
                              "Max_Withdrawals_mgd" = as.numeric(max.withdrawals.mgd))
df_withdrawals <- df_withdrawals %>%
  mutate(Water_System_Name = !!water.system.name,
         PWSID = !!pwsid,
         Ownership = !!ownership,
         Date = my(paste(Month,"-",Year)))

#5
ggplot(df_withdrawals,aes(x=Date,y=Max_Withdrawals_mgd)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2021 Water usage data for",water.system.name),
       subtitle = paste(ownership, pwsid),
       y="Withdrawal (MGD)",
       x="Date")
```

```
## `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 (pwsid) scraped**.

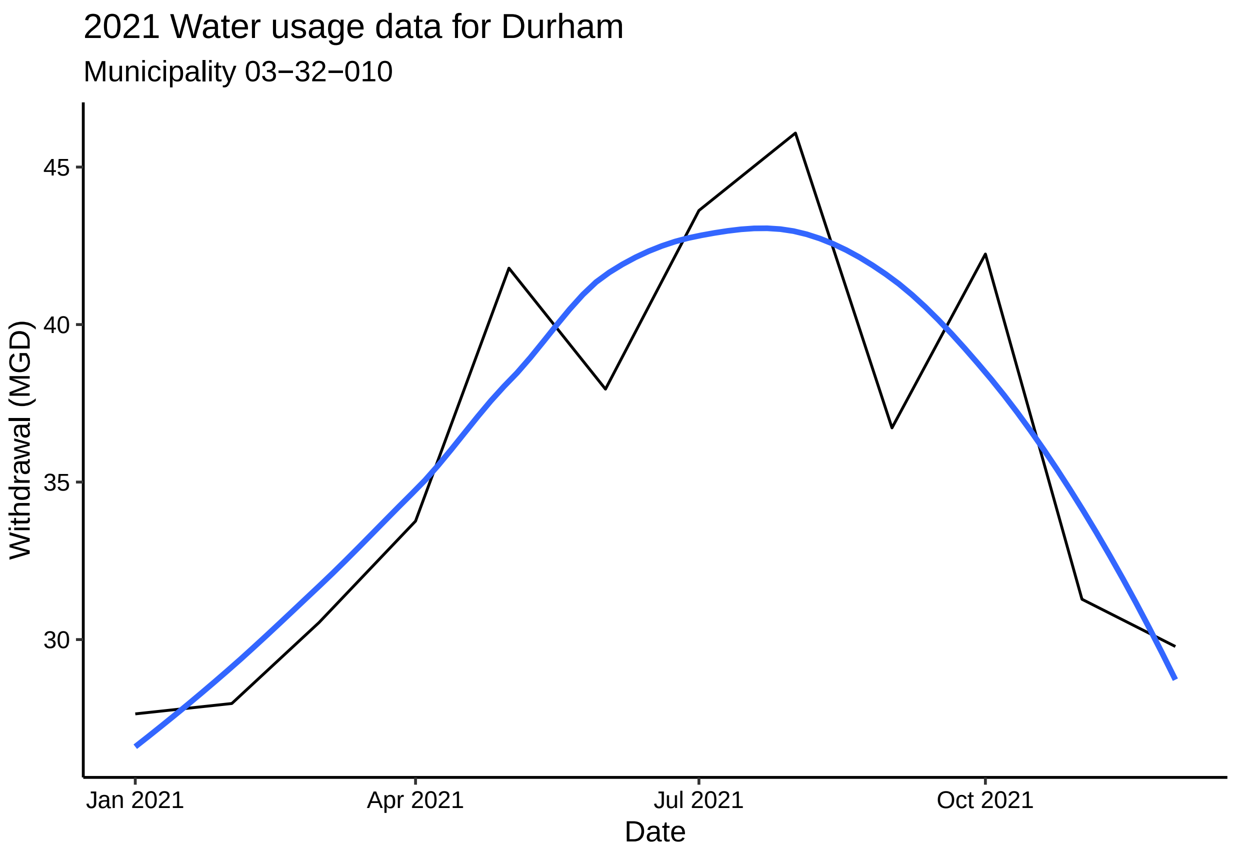
```
#6.
the_base_url <- 'https://www.ncwater.org/WUDC/app/LWSP/report.php?'
the_pwsid <- '03-32-010'
```

```
the_year <- 2021
the_scrape_url <- paste0(the_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=2021"
```

```
the_website <- read_html(the_scrape_url)

the_water.system.name_tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'
the_pwsid_tag <- 'td tr:nth-child(1) td:nth-child(5)'
the_ownership_tag <- 'div+ table tr:nth-child(2) td:nth-child(4)'
the_max.withdrawals.mgd_tag <- 'th~ td+ td'

water.system.name <- the_website %>% html_nodes(the_water.system.name_tag) %>% html_text()
pwsid <- the_website %>%   html_nodes(the_pwsid_tag) %>%  html_text()
ownership <- the_website %>% html_nodes(the_ownership_tag) %>% html_text()
max.withdrawals.mgd <- the_website %>% html_nodes(the_max.withdrawals.mgd_tag) %>% html_text()

df_withdrawals <- data.frame("Month" = c("01", "05", "09", "02", "06", "10",
                                         "03", "07", "11", "04", "08", "12"),
                             "Year" = rep(the_year,12),
                             "Max_Withdrawals_mgd" = as.numeric(max.withdrawals.mgd))
df_withdrawals <- df_withdrawals %>%
  mutate(Water_System_Name = !!water.system.name,
         PWSID = !!pwsid,
         Ownership = !!ownership,
         Date = my(paste(Month,"-",Year)))

ggplot(df_withdrawals,aes(x=Date,y=Max_Withdrawals_mgd)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste(the_year, "Water usage data for",water.system.name),
       subtitle = paste(ownership, pwsid),
       y="Withdrawal (MGD)",
       x="Date")
```

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

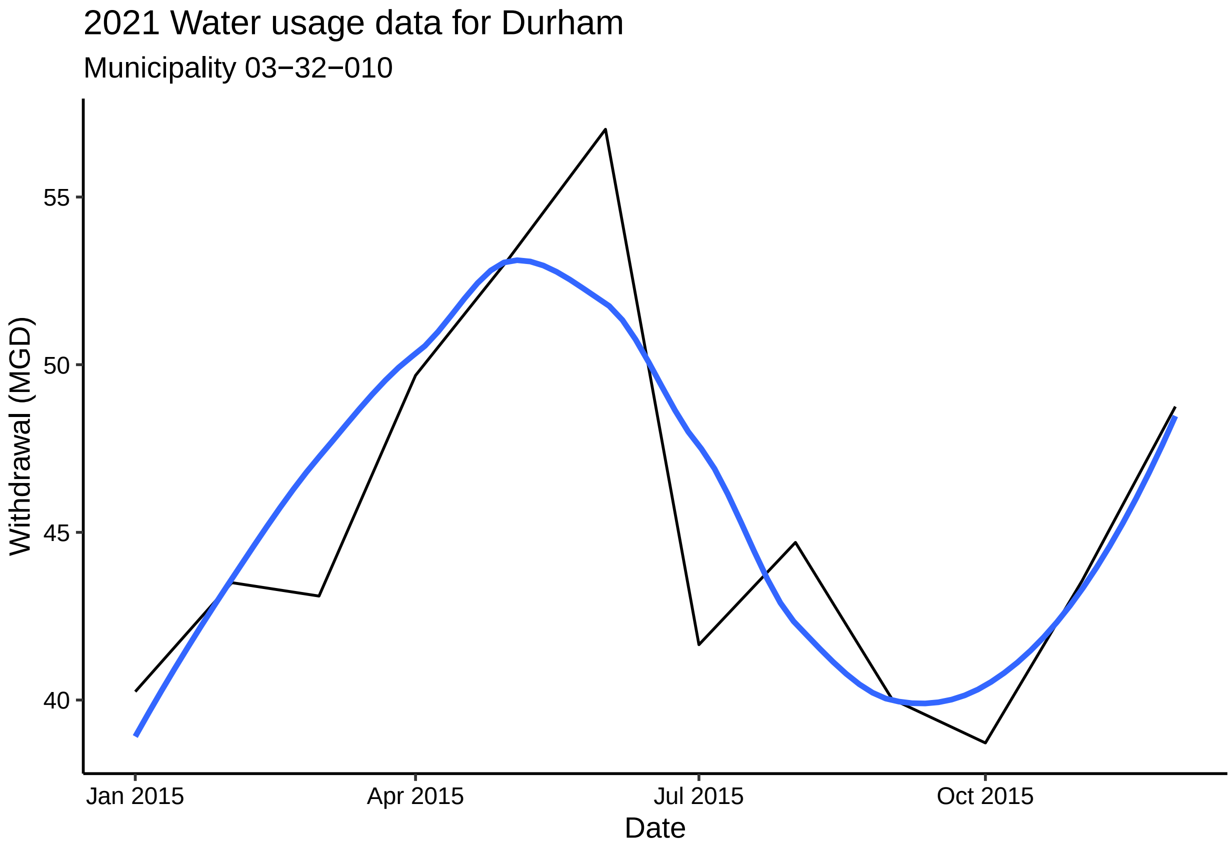


```
#function
scrape.it <- function(the_year, pwsid){
  the_website <- read_html(paste0(the_base_url, 'pwsid=', pwsid, '&year=', the_year))

  the_water.system.name_tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'
  the_pwsid_tag <- 'td tr:nth-child(1) td:nth-child(5)'
  the_ownership_tag <- 'div+ table tr:nth-child(2) td:nth-child(4)'
  the_max.withdrawals.mgd_tag <- 'th~ td+ td'

  water.system.name <- the_website %>% html_nodes(the_water.system.name_tag) %>% html_text()
  pwsid <- the_website %>%   html_nodes(the_pwsid_tag) %>%  html_text()
  ownership <- the_website %>% html_nodes(the_ownership_tag) %>% html_text()
  max.withdrawals.mgd <- the_website %>% html_nodes(the_max.withdrawals.mgd_tag) %>% html_text()

  df_withdrawals <- data.frame("Month" = c("01", "05", "09", "02", "06", "10",
                                           "03", "07", "11", "04", "08", "12"),
                               "Year" = rep(the_year,12),
                               "Max_Withdrawals_mgd" = as.numeric(max.withdrawals.mgd))

  df_withdrawals <- df_withdrawals %>%
    mutate(Water_System_Name = !!water.system.name,
           PWSID = !!pwsid,
           Ownership = !!ownership,
           Date = my(paste(Month,"-",Year)))

  ggplot(df_withdrawals,aes(x=Date,y=Max_Withdrawals_mgd)) +
    geom_line() +
    geom_smooth(method="loess",se=FALSE) +
    labs(title = paste(the_year, "Water usage data for",water.system.name),
         subtitle = paste(ownership, pwsid),
         y="Withdrawal (MGD)",
         x="Date")
  return(df_withdrawals)
}
```

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_2015Durham <- scrape.it(2015,'03-32-010')
view(df_2015Durham)

ggplot(df_2015Durham,aes(x=Date,y=Max_Withdrawals_mgd)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste(the_year, "Water usage data for",water.system.name),
       subtitle = paste(ownership, pwsid),
       y="Withdrawal (MGD)",
       x="Date")
```

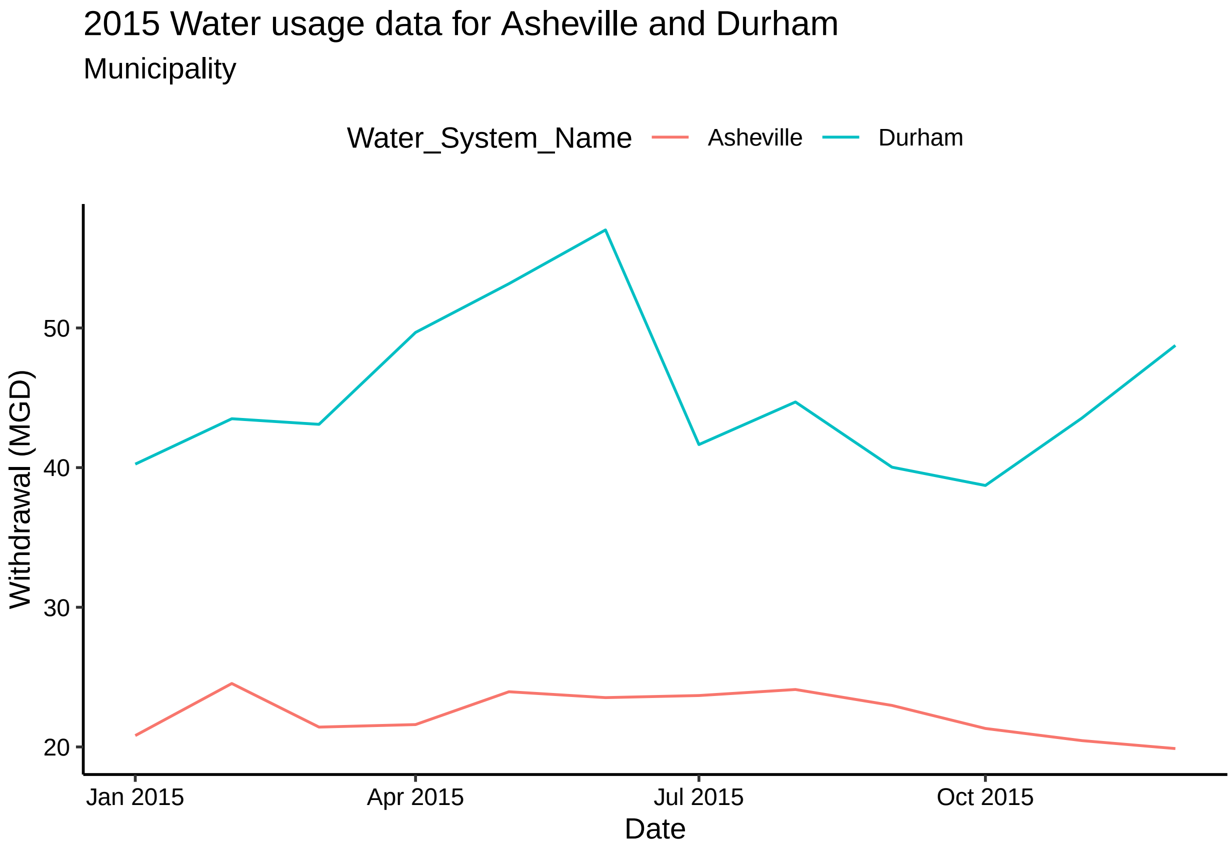
```
## `geom_smooth()` using formula 'y ~ x'
```



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 Asheville's to Durham's water withdrawals.

```
#8
df_2015Asheville <- scrape.it(2015,'01-11-010')
view(df_2015Asheville)

df_combined <- rbind(df_2015Asheville, df_2015Durham) %>%
  group_by(Water_System_Name)

ggplot(df_combined,aes(x=Date,y=Max_Withdrawals_mgd, color=Water_System_Name)) +
  geom_line() +
  labs(title = paste("2015 Water usage data for Asheville and Durham"),
       subtitle = paste(ownership),
       y="Withdrawal (MGD)",
       x="Date")
```



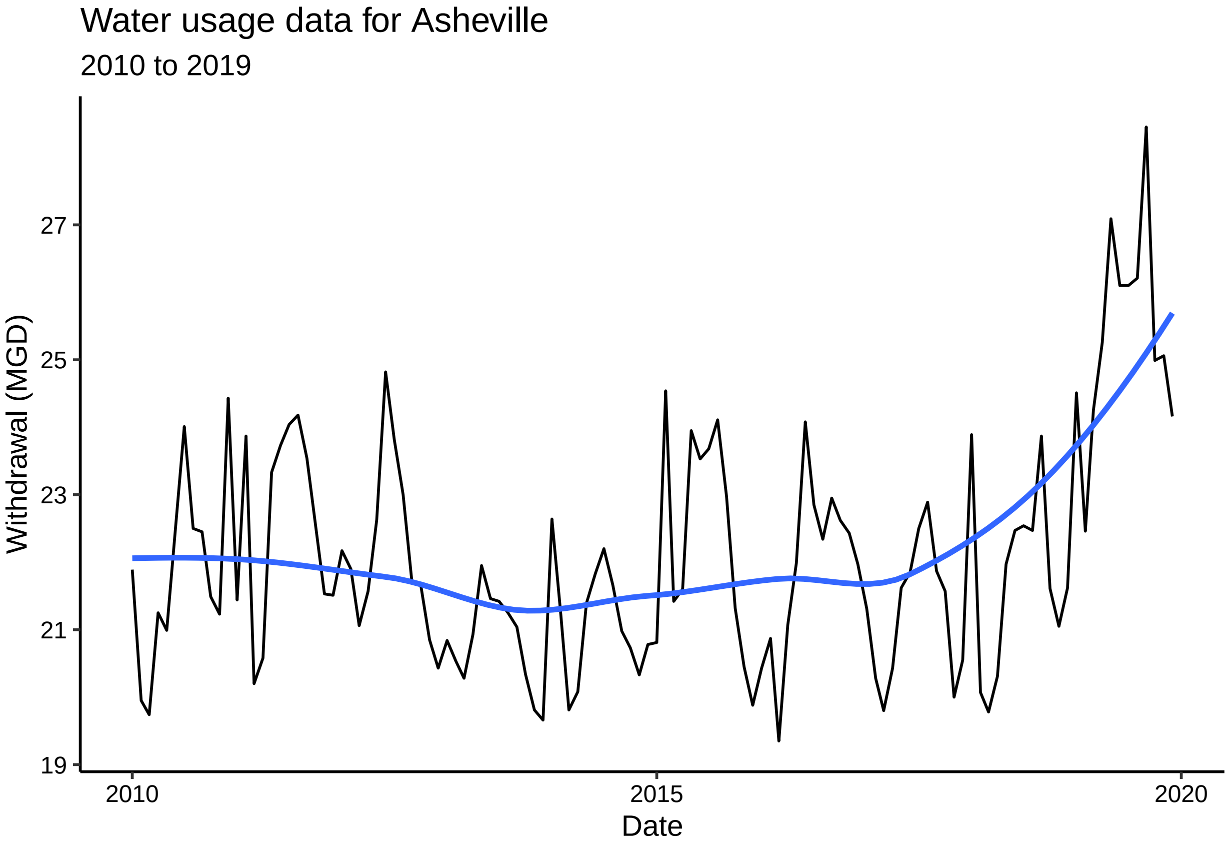
9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

   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(2010:2019)
my_pwsid = '01-11-010'
```

```
the_dfs <- map(the_years,scrape.it,pwsid=my_pwsid)

the_df <- bind_rows(the_dfs)

ggplot(the_df,aes(x=Date,y=Max_Withdrawals_mgd)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("Water usage data for Asheville"),
       subtitle = paste("2010 to 2019"),
       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? Answer: Just by looking at the plot, there is a positive (increasing) trend in water usage over time in Asheville.