Assignment 09: Data Scraping

Sashoy Milton

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() #Check working directory
```

```
## [1] "C:/Users/sasho/Desktop/Environ Data Analytics/Env872 Workspace/EDA-Fall2022_SM/Assignments"
```

```
#Load packages

library(tidyverse)
library(rvest)
```

```
## Warning: package 'rvest' was built under R version 4.2.2
```

```
library(ggplot2)
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.2.2
```

```
## Warning: package 'timechange' was built under R version 4.2.2
```

```
library(dplyr)

# Set theme

my_theme <- theme_bw(base_size = 12) + theme(axis.text = element_text(color = "black"),
legend.position = "top", legend.justification = "center") +
theme(plot.title = element_text(hjust = 0.5))
theme_set(my_theme)
```

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

durham_muncipal <-
read_html('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2021')

durham_muncipal
```

```
## {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".

```r
#3 Scraped and assigned values

water.system.name <- durham_muncipal %>%
  html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>% html_text()

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

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

max.withdrawals.mgd <- durham_muncipal %>%
  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

```r
#4 Create data-frame

local_water_supply.2021 <- data.frame("Month" = c("Jan", "May", "Sept", "Feb", "Jun",
                                      "Oct", "Mar", "Jul", "Nov", "Apr", "Aug","Dec"),
                              "Year" = rep(2021,12),
                              "Max_Withdrawals_mgd" = as.numeric(max.withdrawals.mgd)) %>%
  mutate(Water.System.Name = !!water.system.name,
         PSWID = !!pswid,
         Ownership = !!ownership,
         Date = my(paste(Month,"-",Year)))


local_water_supply.2021 #Check dataset
```

```
##    Month Year Max_Withdrawals_mgd Water.System.Name     PSWID    Ownership
## 1    Jan 2021               27.64            Durham 03-32-010 Municipality
## 2    May 2021               41.79            Durham 03-32-010 Municipality
## 3   Sept 2021               36.72            Durham 03-32-010 Municipality
## 4    Feb 2021               27.97            Durham 03-32-010 Municipality
## 5    Jun 2021               37.95            Durham 03-32-010 Municipality
## 6    Oct 2021               42.24            Durham 03-32-010 Municipality
## 7    Mar 2021               30.54            Durham 03-32-010 Municipality
## 8    Jul 2021               43.62            Durham 03-32-010 Municipality
## 9    Nov 2021               31.28            Durham 03-32-010 Municipality
## 10   Apr 2021               33.76            Durham 03-32-010 Municipality
```

```
## 11   Aug 2021               46.08            Durham 03-32-010 Municipality
## 12   Dec 2021               29.78            Durham 03-32-010 Municipality
##          Date
## 1  2021-01-01
## 2  2021-05-01
## 3  2021-09-01
## 4  2021-02-01
## 5  2021-06-01
## 6  2021-10-01
## 7  2021-03-01
## 8  2021-07-01
## 9  2021-11-01
## 10 2021-04-01
## 11 2021-08-01
## 12 2021-12-01
```

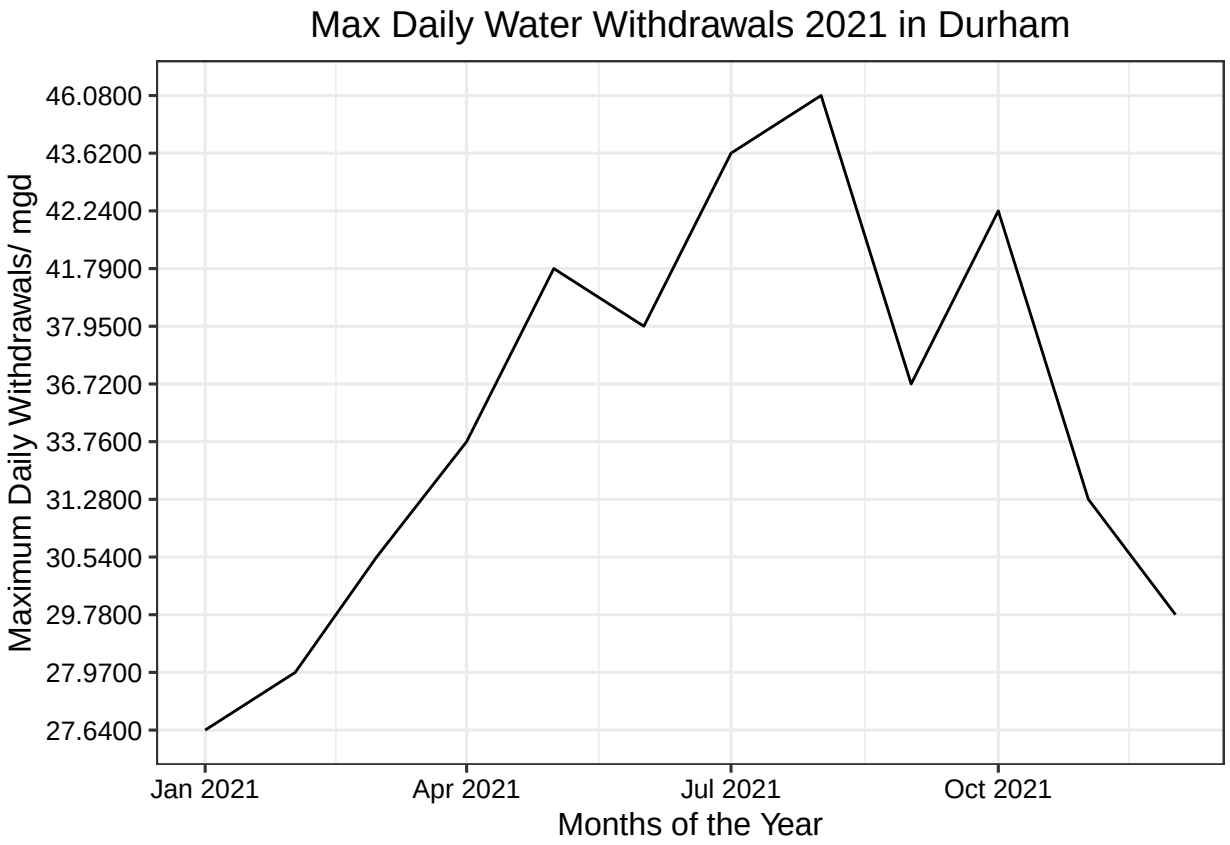
```
#Check variable type

class(local_water_supply.2021$Date)
```

```
## [1] "Date"
```

```
class(local_water_supply.2021$Max_Withdrawals_mgd)
```

```
## [1] "numeric"
```

```
#5. Create line plot of maximum daily withdrawals

ggplot(local_water_supply.2021, aes(x = Date, y = max.withdrawals.mgd, group = 1)) +
  geom_line() +
  ylab ("Maximum Daily Withdrawals/ mgd") +
  xlab ("Months of the Year") +
  labs(title ="Max Daily Water Withdrawals 2021 in Durham")
```



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.


scrape_website <- function(the_pswid,the_year){


#Retrieve the website contents

the_website <-read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',
                               the_pswid,'&year=',the_year))

#Set the element address variables
  water_system_name_tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'
  pswid_tag <- 'td tr:nth-child(1) td:nth-child(5)'
  ownership_tag <- 'div+ table tr:nth-child(2) td:nth-child(4)'
  max_withdrawal_tag <- 'th~ td+ td'

#Scrape the data items
water.system.name <- the_website %>% html_nodes(water_system_name_tag) %>% html_text()
pswid <-the_website %>% html_nodes(pswid_tag) %>% html_text()
ownership <- the_website %>% html_nodes(ownership_tag) %>% html_text()
```

```
max.withdrawals.mgd <- the_website %>% html_nodes(max_withdrawal_tag) %>% html_text()

#Convert to a dataframe
 water_supply <- data.frame("Month" = c("Jan", "May", "Sept", "Feb", "Jun",
                                        "Oct", "Mar", "Jul", "Nov", "Apr", "Aug","Dec"),
                            "Year" = rep(the_year,12),
                            "Max_Withdrawals_mgd" = as.numeric(max.withdrawals.mgd)) %>%
 mutate(Water.System.Name = !!water.system.name,
         PSWID = !!pswid,
         Ownership = !!ownership,
         Date = my(paste(Month,"-",Year)))

 Sys.sleep(1)

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

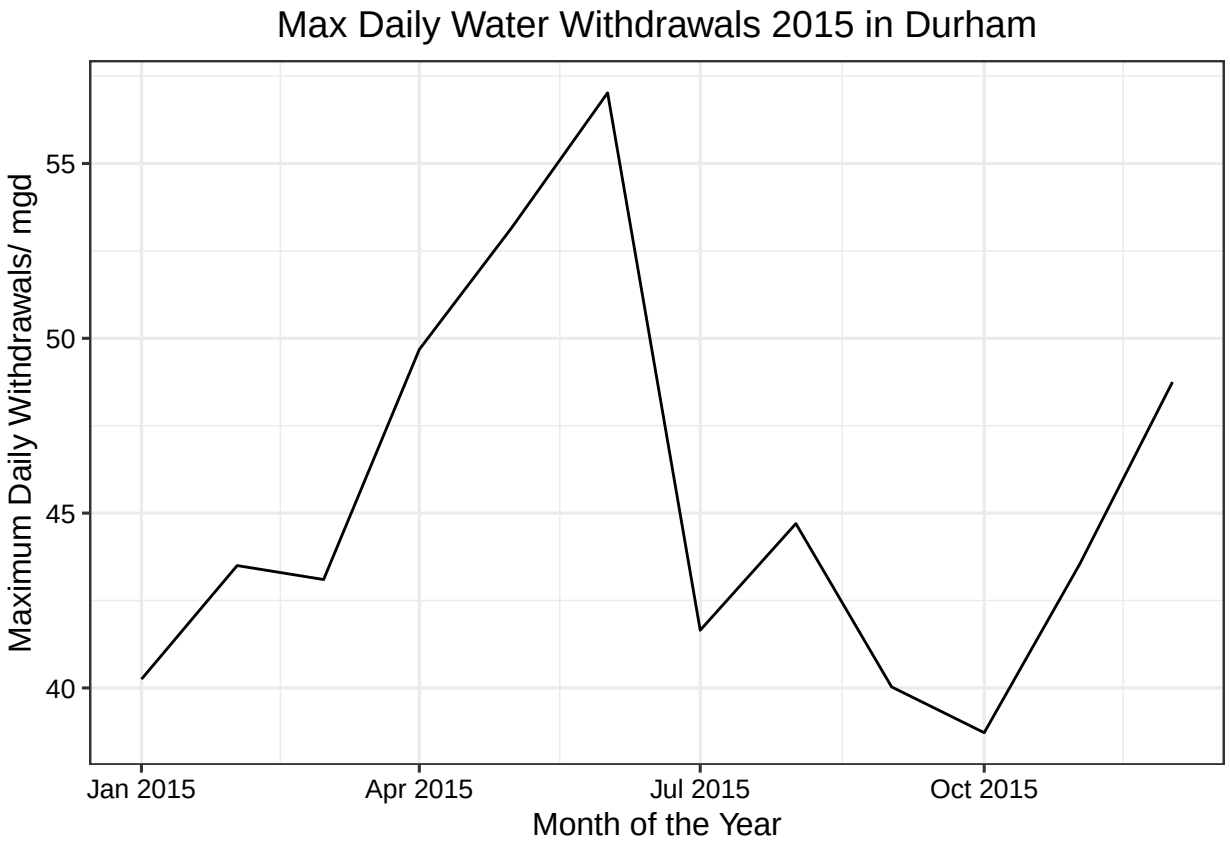
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID=‘03-32-010’) for each month in 2015

```
#7

# Extract Max Daily Withdrawals for Durham

local_water_supply.2015 <- scrape_website('03-32-010',2015)
local_water_supply.2015 # View data set
```

```
##    Month Year Max_Withdrawals_mgd Water.System.Name     PSWID    Ownership
## 1    Jan 2015               40.25            Durham 03-32-010 Municipality
## 2    May 2015               53.17            Durham 03-32-010 Municipality
## 3   Sept 2015               40.03            Durham 03-32-010 Municipality
## 4    Feb 2015               43.50            Durham 03-32-010 Municipality
## 5    Jun 2015               57.02            Durham 03-32-010 Municipality
## 6    Oct 2015               38.72            Durham 03-32-010 Municipality
## 7    Mar 2015               43.10            Durham 03-32-010 Municipality
## 8    Jul 2015               41.65            Durham 03-32-010 Municipality
## 9    Nov 2015               43.55            Durham 03-32-010 Municipality
## 10   Apr 2015               49.68            Durham 03-32-010 Municipality
## 11   Aug 2015               44.70            Durham 03-32-010 Municipality
## 12   Dec 2015               48.75            Durham 03-32-010 Municipality
##          Date
## 1  2015-01-01
## 2  2015-05-01
## 3  2015-09-01
## 4  2015-02-01
## 5  2015-06-01
## 6  2015-10-01
## 7  2015-03-01
## 8  2015-07-01
## 9  2015-11-01
## 10 2015-04-01
## 11 2015-08-01
```

```
## 12 2015-12-01
```

```
# Line plot of maximum daily withdrawals

ggplot(local_water_supply.2015, aes(x = Date, y = Max_Withdrawals_mgd, group = 1)) +
  geom_line() +
  ylab ("Maximum Daily Withdrawals/ mgd") +
  xlab ("Month of the Year") +
  labs(title = ("Max Daily Water Withdrawals 2015 in Durham"))
```



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
##Extract Asheville data

local_water_supply_2015_Asheville <- scrape_website('01-11-010',2015)
local_water_supply_2015_Asheville #View data set
```

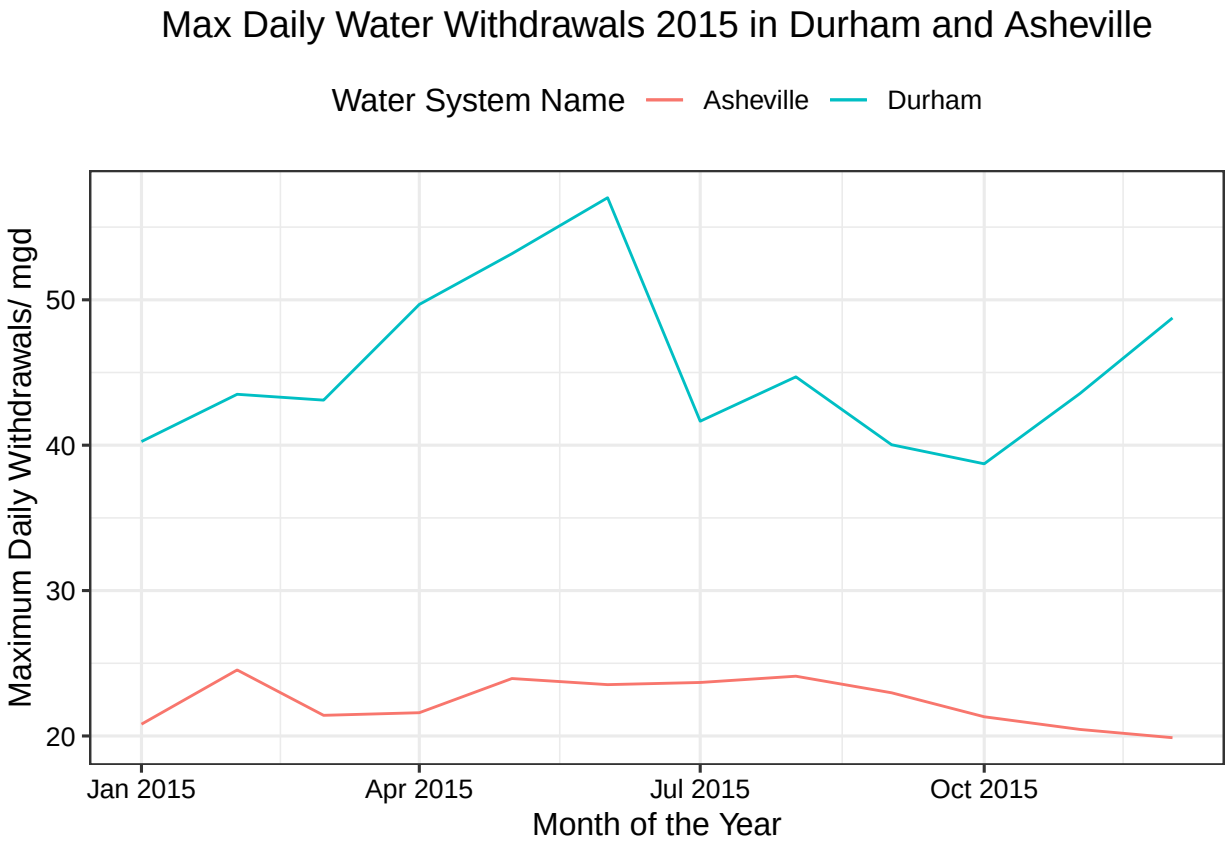
```
##    Month Year Max_Withdrawals_mgd Water.System.Name     PSWID    Ownership
## 1    Jan 2015               20.81         Asheville 01-11-010 Municipality
## 2    May 2015               23.95         Asheville 01-11-010 Municipality
## 3   Sept 2015               22.97         Asheville 01-11-010 Municipality
```

```
## 4    Feb 2015              24.54         Asheville 01-11-010 Municipality
## 5    Jun 2015              23.53         Asheville 01-11-010 Municipality
## 6    Oct 2015              21.32         Asheville 01-11-010 Municipality
## 7    Mar 2015              21.42         Asheville 01-11-010 Municipality
## 8    Jul 2015              23.68         Asheville 01-11-010 Municipality
## 9    Nov 2015              20.45         Asheville 01-11-010 Municipality
## 10   Apr 2015              21.60         Asheville 01-11-010 Municipality
## 11   Aug 2015              24.11         Asheville 01-11-010 Municipality
## 12   Dec 2015              19.88         Asheville 01-11-010 Municipality
##          Date
## 1  2015-01-01
## 2  2015-05-01
## 3  2015-09-01
## 4  2015-02-01
## 5  2015-06-01
## 6  2015-10-01
## 7  2015-03-01
## 8  2015-07-01
## 9  2015-11-01
## 10 2015-04-01
## 11 2015-08-01
## 12 2015-12-01
```

```
colnames(local_water_supply_2015_Asheville)
```

```
## [1] "Month"               "Year"                "Max_Withdrawals_mgd"
## [4] "Water.System.Name"   "PSWID"               "Ownership"
## [7] "Date"
```

```
#Create a plot that compares Asheville's to Durham's water withdrawals

ggplot(local_water_supply_2015_Asheville, aes(x=Date, y=Max_Withdrawals_mgd,
                                    group = 1, color = Water.System.Name))+
  geom_line() +
  geom_line(data = local_water_supply.2015, aes(x = Date, y = Max_Withdrawals_mgd,
                                    group = 1, color = water.system.name)) +
    ylab ("Maximum Daily Withdrawals/ mgd") +
  xlab ("Month of the Year") +
  labs(title = ("Max Daily Water Withdrawals 2015 in Durham and Asheville"),
       color = "Water System Name")
```



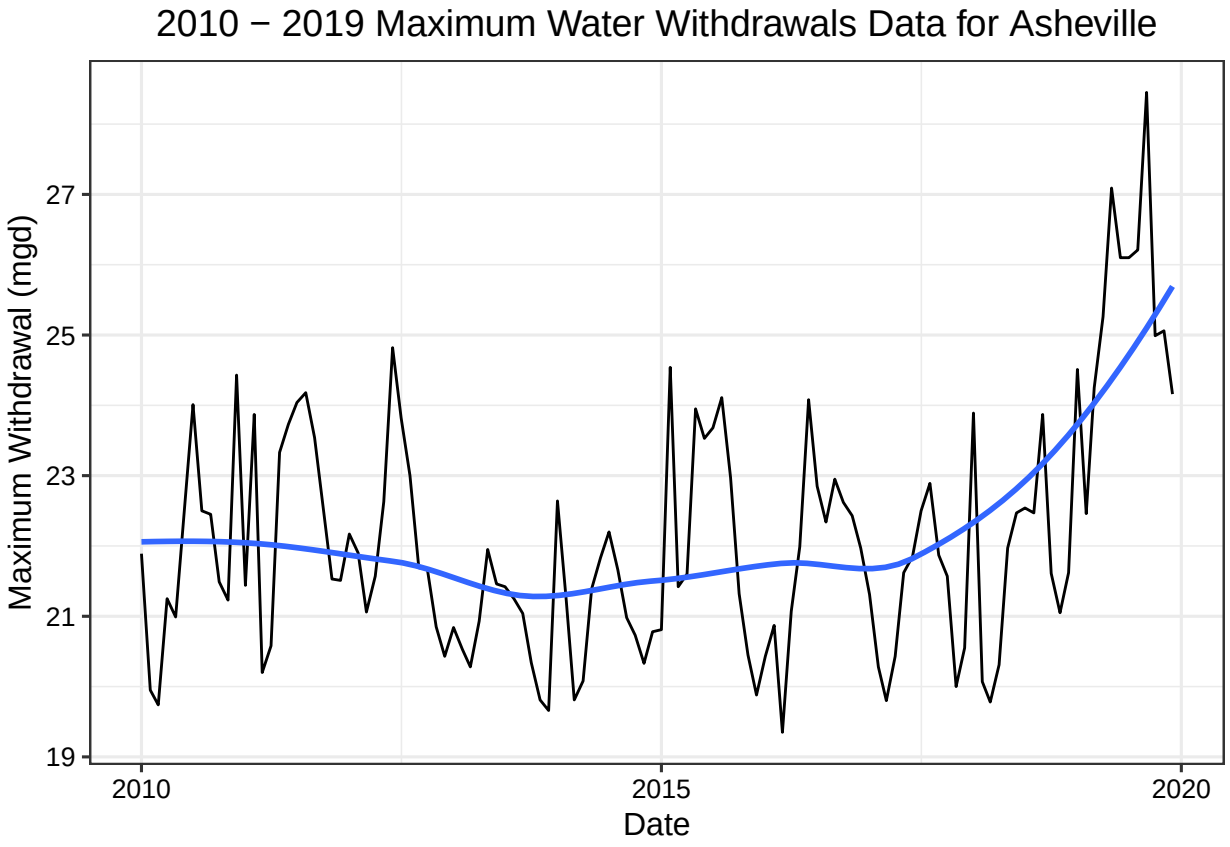
9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

   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

#Set the inputs to scrape years 2015 to 2020 for the Asheville site "01-11-010"
the_years = rep(2010:2019)
my_facility = '01-11-010'

# Map function

the_dfs <- map2(my_facility,the_years,scrape_website)

#Conflate the returned dataframes into a single dataframe
the_df <- bind_rows(the_dfs)

#Plot Asheville's max daily withdrawal by months for the years 2010 thru 2019
ggplot(the_df,aes(x=Date,y=Max_Withdrawals_mgd)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2010 - 2019 Maximum Water Withdrawals Data for Asheville"),
```

```
       y= "Maximum 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, Asheville does have a trend in water usage over time. The maximum water usuage was relatively constant at about 22 mgd from the year 2010 to around 2017, when a sharp increase in maximum withdrawal was seen up to about 26 mgd.