Assignment 10: Data Scraping

Gary Alvarez Mejia

OVERVIEW

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

Directions

1. Rename this file `<FirstLast>_A10_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 your code is tidy; use line breaks to ensure your code fits in the knitted output.
5. Be sure to **answer the questions** in this assignment document.
6. When you have completed the assignment, **Knit** the text and code into a single PDF file.

Set up

1. Set up your session:

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

```
#1
library(tidyverse)
library(lubridate)
library(here); here()
```

```
## [1] "/home/guest/R/EDA_Spring2024"
```

```
#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 2022 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=2022

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

```
#2
#Fetch the web resources from the URL
webpage <- read_html(
  'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2022')
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
- PWSID
- Ownership
- From the "3. Water Supply Sources" section:
- Maximum Day 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)".

```
#3
water_system_name <- webpage %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()
water_system_name
```

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

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

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

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

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

```
max_day_use <- webpage %>%
  html_nodes("th~ td+ td") %>%
  html_text()
max_day_use
```

```
##  [1] "36.1000" "43.4200" "52.4900" "30.5000" "42.5900" "34.8800" "39.9100"
##  [8] "43.3200" "32.5300" "34.6600" "41.8000" "37.5300"
```

```
months <- webpage %>%
  html_nodes(".fancy-table:nth-child(31) tr+ tr th") %>%
  html_text()
months
```

```
##  [1] "Jan" "May" "Sep" "Feb" "Jun" "Oct" "Mar" "Jul" "Nov" "Apr" "Aug" "Dec"
```

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.

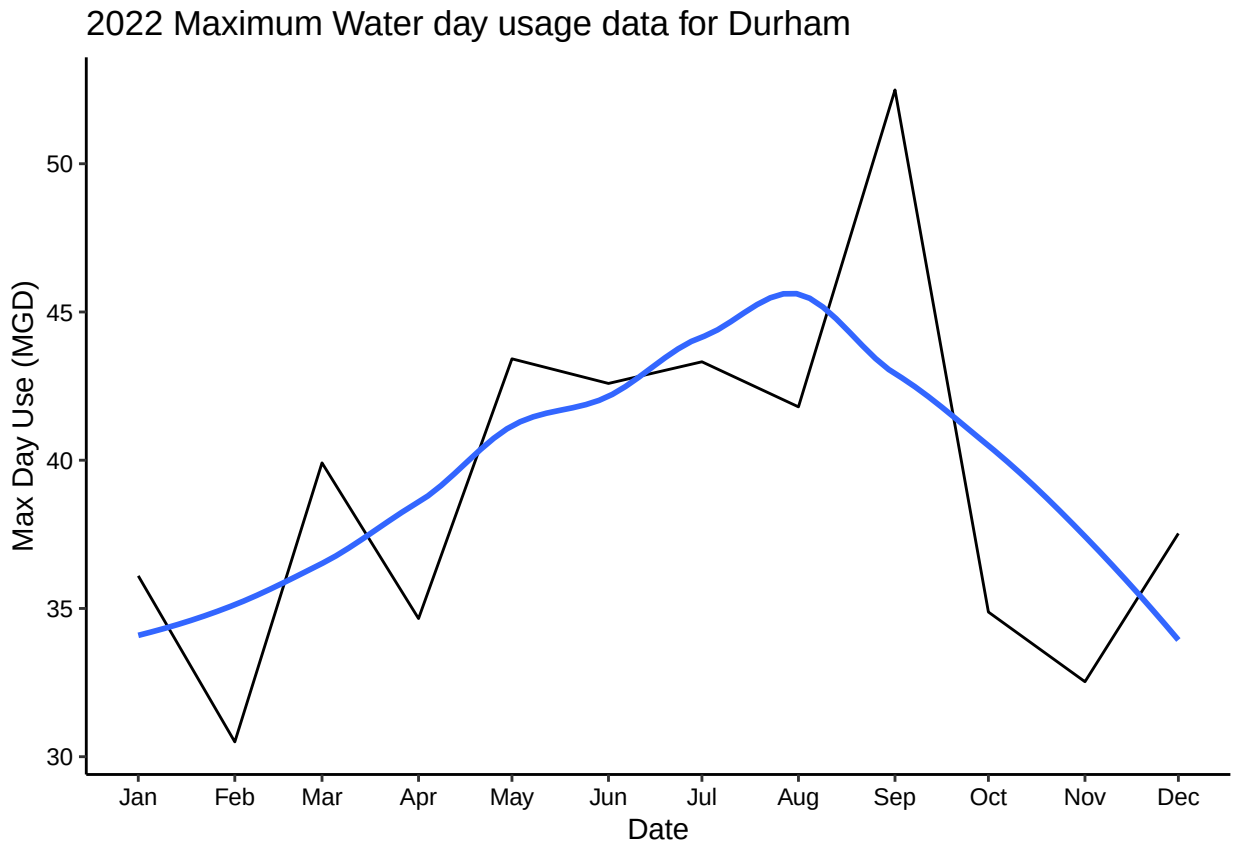
   NOTE: It's likely you won't be able to scrape the monthly widthrawal data in chronological order. You can overcome this by creating a month column manually assigning values in the order the data are scraped: "Jan", "May", "Sept", "Feb", etc... Or, you could scrape month values from the web page...

5. Create a line plot of the maximum daily withdrawals across the months for 2022

```
#4
df_withdrawals <- data.frame("Ownership" = ownership,
                             "PWSID" = PWSID,
                             "Water System Name" = water_system_name,
                             "Month" = rep(c(months)),
                             "Year" = rep(2022, each = 12),
                             "Max Day Use" = as.numeric(max_day_use))

df_withdrawals <- df_withdrawals %>%
  mutate(Date = my(paste(Month,"-",Year)))

#5
ggplot(df_withdrawals,aes(x=Date,y=Max.Day.Use)) +
  geom_line() +
```

```
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2022 Maximum Water day usage data for",water_system_name),
       y="Max Day Use (MGD)",
       x="Date") +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")
```

```
## `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.
scrape.it <- function(the_year, the_PWSID){

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

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_day_use <- webpage %>%
  html_nodes("th~ td+ td") %>%
  html_text()

df_withdrawals_func <- data.frame("Ownership" = ownership,
                                  "PWSID" = PWSID,
                                  "City" = water_system_name,
                                  "Month" = months,
                                  "Year" = rep(the_year, each = 12),
                                  "Max Day Use" = as.numeric(max_day_use))

return(df_withdrawals_func)
}

#Trying out the function
scrape.it(2014, "01-11-010")
```

```
##       Ownership     PWSID      City Month Year Max.Day.Use
## 1  Municipality 01-11-010 Asheville   Jan 2014       22.64
## 2  Municipality 01-11-010 Asheville   May 2014       21.39
## 3  Municipality 01-11-010 Asheville   Sep 2014       20.98
## 4  Municipality 01-11-010 Asheville   Feb 2014       21.22
## 5  Municipality 01-11-010 Asheville   Jun 2014       21.83
## 6  Municipality 01-11-010 Asheville   Oct 2014       20.73
## 7  Municipality 01-11-010 Asheville   Mar 2014       19.81
## 8  Municipality 01-11-010 Asheville   Jul 2014       22.20
## 9  Municipality 01-11-010 Asheville   Nov 2014       20.33
## 10 Municipality 01-11-010 Asheville   Apr 2014       20.08
## 11 Municipality 01-11-010 Asheville   Aug 2014       21.66
## 12 Municipality 01-11-010 Asheville   Dec 2014       20.78
```

```
df_withdrawals_q6 <- scrape.it(2014, "01-11-010")
```

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

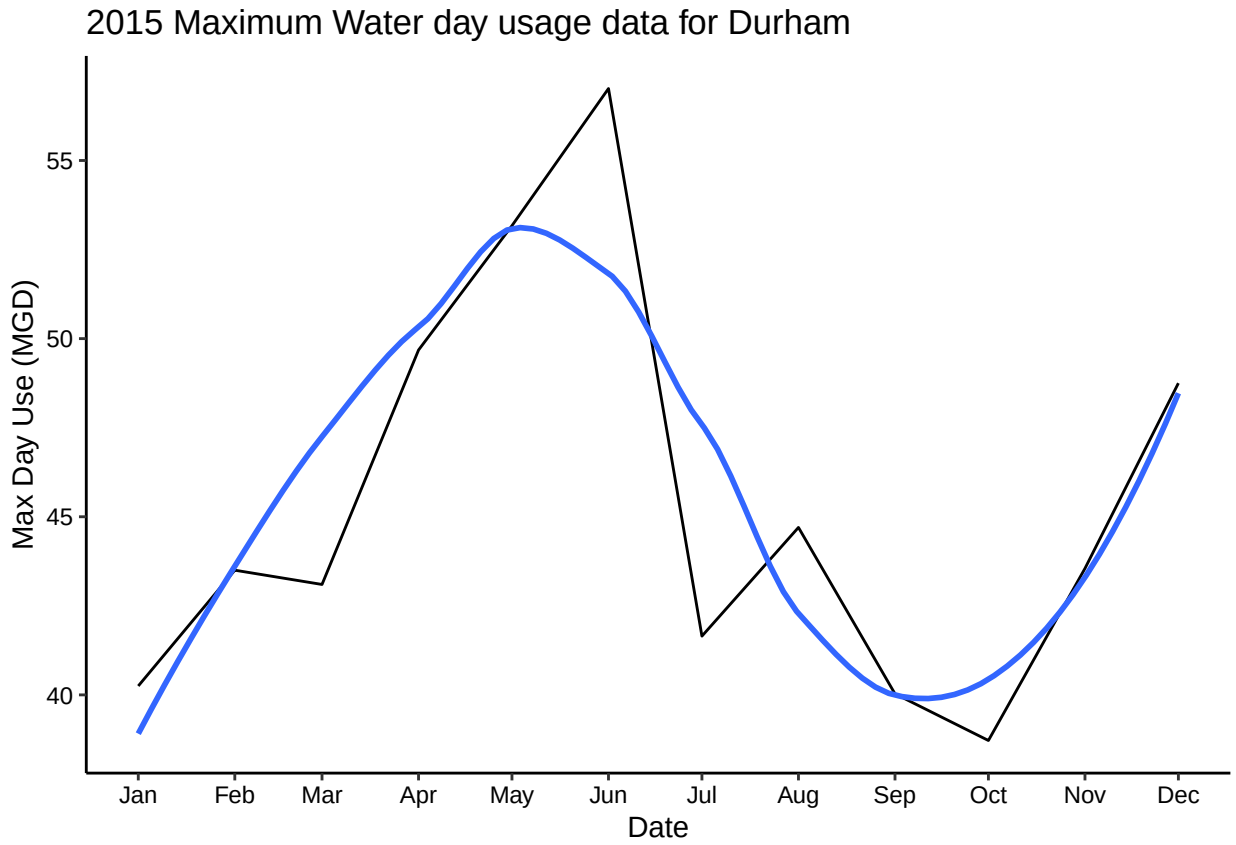
```
#7
scrape.it(2015, "03-32-010")
```

```
##       Ownership     PWSID   City Month Year Max.Day.Use
## 1  Municipality 03-32-010 Durham   Jan 2015       40.25
## 2  Municipality 03-32-010 Durham   May 2015       53.17
## 3  Municipality 03-32-010 Durham   Sep 2015       40.03
## 4  Municipality 03-32-010 Durham   Feb 2015       43.50
```

```
## 5  Municipality 03-32-010 Durham   Jun 2015       57.02
## 6  Municipality 03-32-010 Durham   Oct 2015       38.72
## 7  Municipality 03-32-010 Durham   Mar 2015       43.10
## 8  Municipality 03-32-010 Durham   Jul 2015       41.65
## 9  Municipality 03-32-010 Durham   Nov 2015       43.55
## 10 Municipality 03-32-010 Durham   Apr 2015       49.68
## 11 Municipality 03-32-010 Durham   Aug 2015       44.70
## 12 Municipality 03-32-010 Durham   Dec 2015       48.75
```

```r
df_withdrawals_q7 <- scrape.it(2015, "03-32-010")

df_withdrawals_q7 <- df_withdrawals_q7 %>%
  mutate(Date = my(paste(Month,"-",Year)))

ggplot(df_withdrawals_q7,aes(x=Date,y=Max.Day.Use)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste(2015, "Maximum Water day usage data for", "Durham"),
       y="Max Day Use (MGD)",
       x="Date") +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")
```

```
## `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

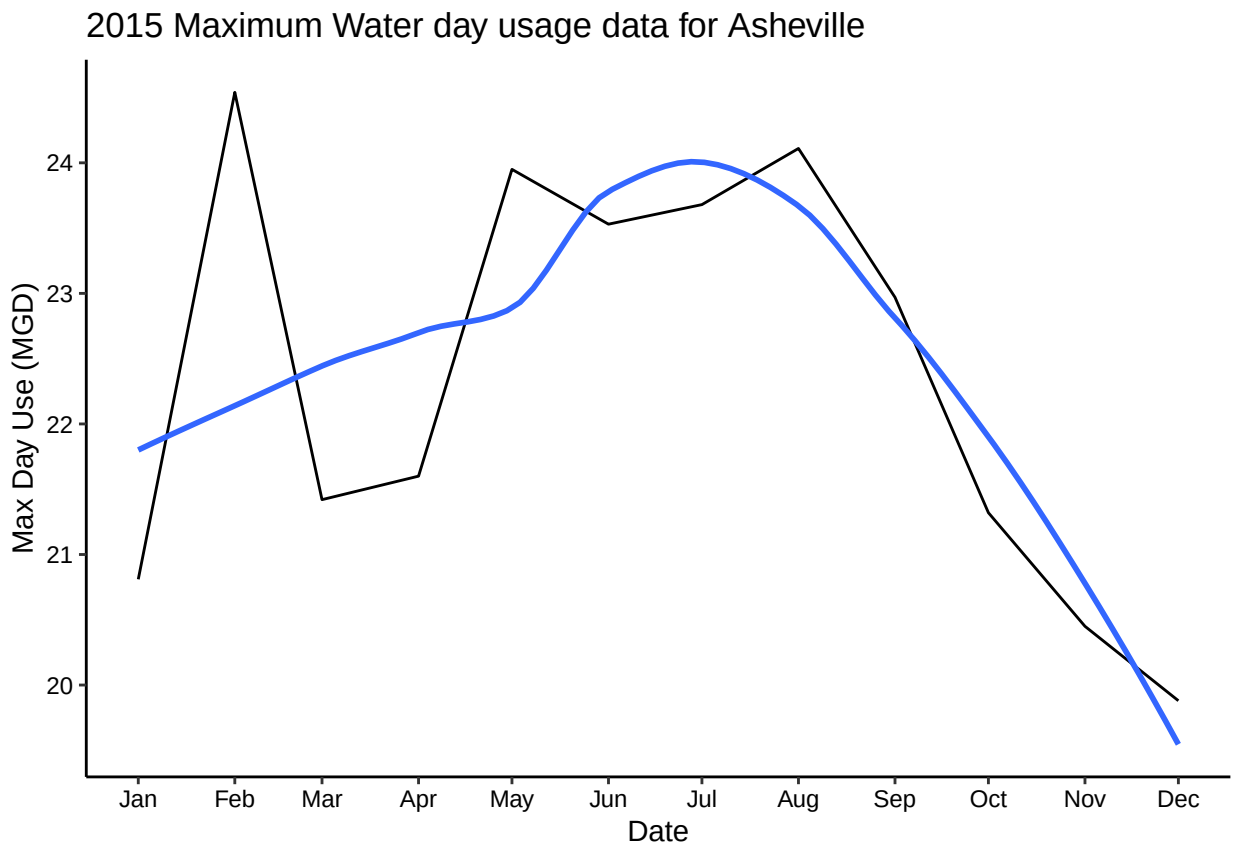
with the Durham data collected above and create a plot that compares Asheville's to Durham's water withdrawals.

```
#8
scrape.it(2015, "01-11-010")
```

```
##       Ownership     PWSID      City Month Year Max.Day.Use
## 1  Municipality 01-11-010 Asheville   Jan 2015       20.81
## 2  Municipality 01-11-010 Asheville   May 2015       23.95
## 3  Municipality 01-11-010 Asheville   Sep 2015       22.97
## 4  Municipality 01-11-010 Asheville   Feb 2015       24.54
## 5  Municipality 01-11-010 Asheville   Jun 2015       23.53
## 6  Municipality 01-11-010 Asheville   Oct 2015       21.32
## 7  Municipality 01-11-010 Asheville   Mar 2015       21.42
## 8  Municipality 01-11-010 Asheville   Jul 2015       23.68
## 9  Municipality 01-11-010 Asheville   Nov 2015       20.45
## 10 Municipality 01-11-010 Asheville   Apr 2015       21.60
## 11 Municipality 01-11-010 Asheville   Aug 2015       24.11
## 12 Municipality 01-11-010 Asheville   Dec 2015       19.88
```

```
df_withdrawals_q8 <- scrape.it(2015, "01-11-010")

df_withdrawals_q8 <- df_withdrawals_q8 %>%
  mutate(Date = my(paste(Month,"-",Year)))

ggplot(df_withdrawals_q8,aes(x=Date,y=Max.Day.Use)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste(2015, "Maximum Water day usage data for", "Asheville"),
       y="Max Day Use (MGD)",
       x="Date") +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")
```

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



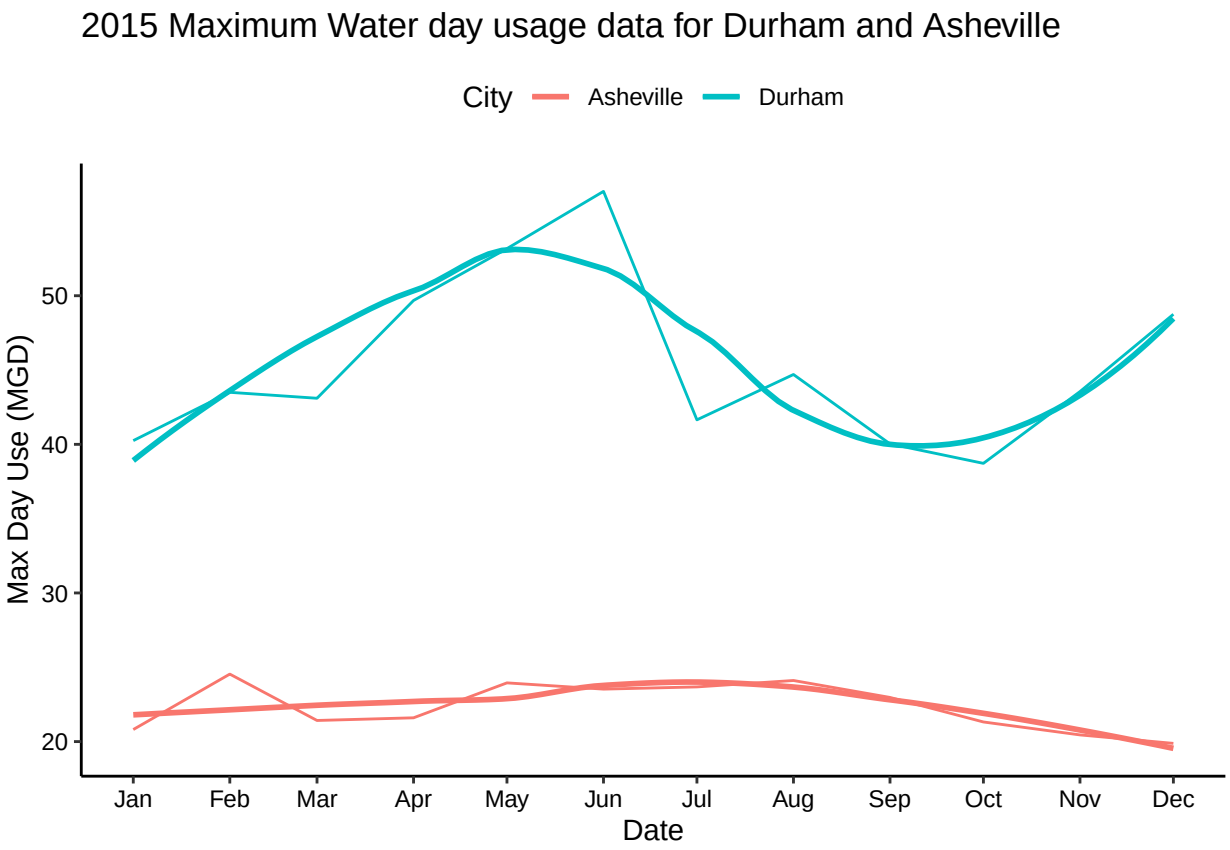

```
df_combined <- bind_rows(df_withdrawals_q7, df_withdrawals_q8)

df_combined
```

```
##       Ownership     PWSID      City Month Year Max.Day.Use       Date
## 1  Municipality 03-32-010    Durham   Jan 2015       40.25 2015-01-01
## 2  Municipality 03-32-010    Durham   May 2015       53.17 2015-05-01
## 3  Municipality 03-32-010    Durham   Sep 2015       40.03 2015-09-01
## 4  Municipality 03-32-010    Durham   Feb 2015       43.50 2015-02-01
## 5  Municipality 03-32-010    Durham   Jun 2015       57.02 2015-06-01
## 6  Municipality 03-32-010    Durham   Oct 2015       38.72 2015-10-01
## 7  Municipality 03-32-010    Durham   Mar 2015       43.10 2015-03-01
## 8  Municipality 03-32-010    Durham   Jul 2015       41.65 2015-07-01
## 9  Municipality 03-32-010    Durham   Nov 2015       43.55 2015-11-01
## 10 Municipality 03-32-010    Durham   Apr 2015       49.68 2015-04-01
## 11 Municipality 03-32-010    Durham   Aug 2015       44.70 2015-08-01
## 12 Municipality 03-32-010    Durham   Dec 2015       48.75 2015-12-01
## 13 Municipality 01-11-010 Asheville   Jan 2015       20.81 2015-01-01
## 14 Municipality 01-11-010 Asheville   May 2015       23.95 2015-05-01
## 15 Municipality 01-11-010 Asheville   Sep 2015       22.97 2015-09-01
## 16 Municipality 01-11-010 Asheville   Feb 2015       24.54 2015-02-01
## 17 Municipality 01-11-010 Asheville   Jun 2015       23.53 2015-06-01
## 18 Municipality 01-11-010 Asheville   Oct 2015       21.32 2015-10-01
## 19 Municipality 01-11-010 Asheville   Mar 2015       21.42 2015-03-01
## 20 Municipality 01-11-010 Asheville   Jul 2015       23.68 2015-07-01
```

```
## 21 Municipality 01-11-010 Asheville   Nov 2015       20.45 2015-11-01
## 22 Municipality 01-11-010 Asheville   Apr 2015       21.60 2015-04-01
## 23 Municipality 01-11-010 Asheville   Aug 2015       24.11 2015-08-01
## 24 Municipality 01-11-010 Asheville   Dec 2015       19.88 2015-12-01
```

```
ggplot(df_combined,aes(x=Date,y=Max.Day.Use, color=City)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste(2015, "Maximum Water day usage data for Durham and Asheville"),
       y="Max Day Use (MGD)",
       x="Date") +
  scale_x_date(date_breaks = "1 month", date_labels = "%b")
```

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

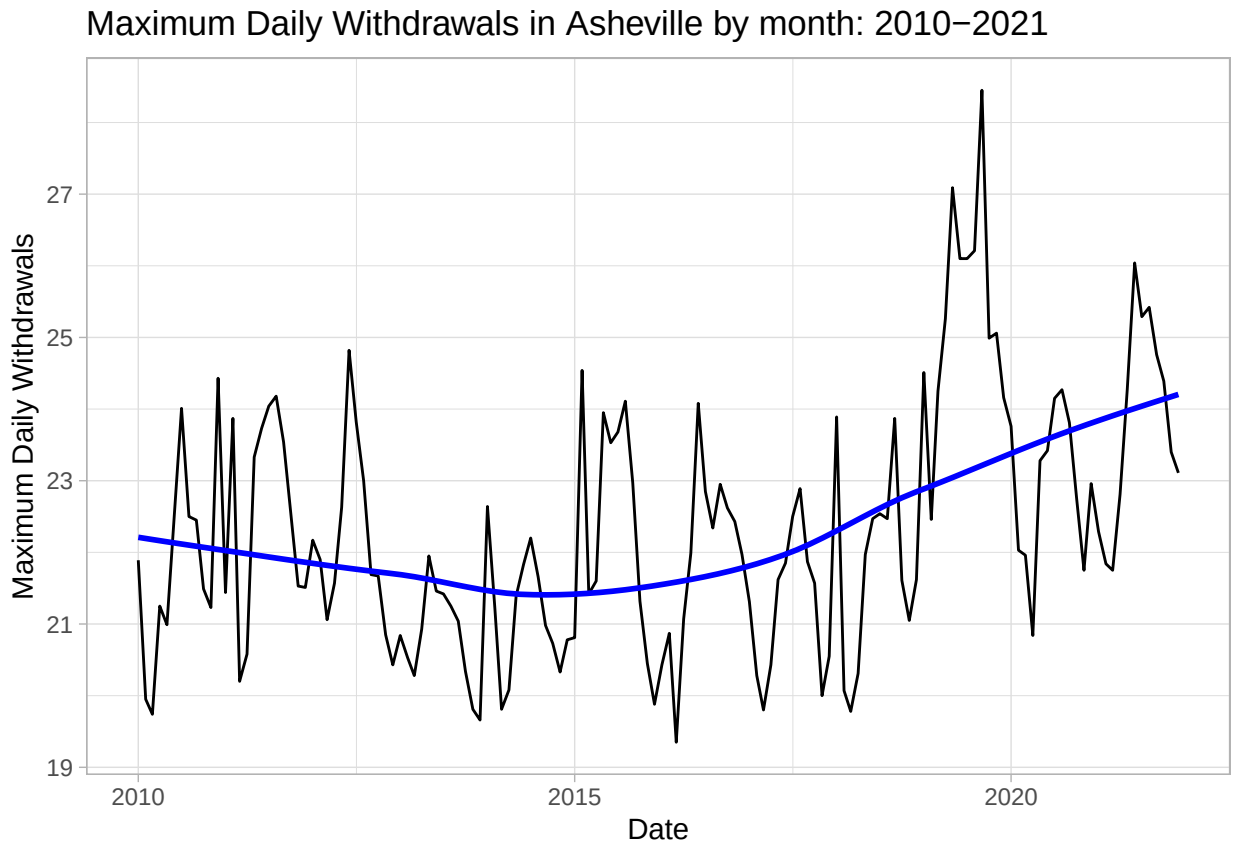


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2021.Add a smoothed line to the plot (method = 'loess').

   TIP: See Section 3.2 in the "10_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.

```r
#9
Asheville_decade <- 2010:2021
Asheville_id <- rep('01-11-010', length(Asheville_decade))

df_Asheville <- map2(Asheville_decade, Asheville_id, scrape.it)
df_Asheville <- bind_rows(df_Asheville)

df_Asheville <- df_Asheville %>%
  mutate(Date = my(paste(Month,"-",Year)))

Asheville_plot <- ggplot(df_Asheville, aes(x=Date,y=Max.Day.Use)) +
  geom_line() +
  geom_smooth(method = 'loess', se = F, color = 'blue') +
  labs(x = "Date", y = "Maximum Daily Withdrawals") +
  theme_light() +
  ggtitle("Maximum Daily Withdrawals in Asheville by month: 2010-2021")
Asheville_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? > Answer: Yes, Asheville has a trend in water usage over time. The plot shows that from 2010 to 2015 the water usage slighltly decreased, but it started increasing significantly after 2015 (and in the 2015-2021 period). >