Assignment 10: Data Scraping

Student Name

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

```
## Warning: package 'rvest' was built under R version 4.2.3
```

```
library(tidyverse)

getwd()
```

```
## [1] "X:/ENV 872 Environmental Data Analytics/Git_codes/EDA-Spring2023"
```

```
my_theme <- theme_minimal() +
  theme(
    legend.position = "bottom",
    legend.title = element_text(colour = "#440000"),
    plot.title = element_text(colour = "blue",
                              hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5),
    axis.title = element_text(colour = "#4169e1")
  )
theme_set(my_theme)
```

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
durham_lwsp.web <- read_html("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2022
```

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), with the first value being "36.1000".

```
#3
water.system.name <- durham_lwsp.web %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()
water.system.name
```

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

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

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

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

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

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

```
##  [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"
```

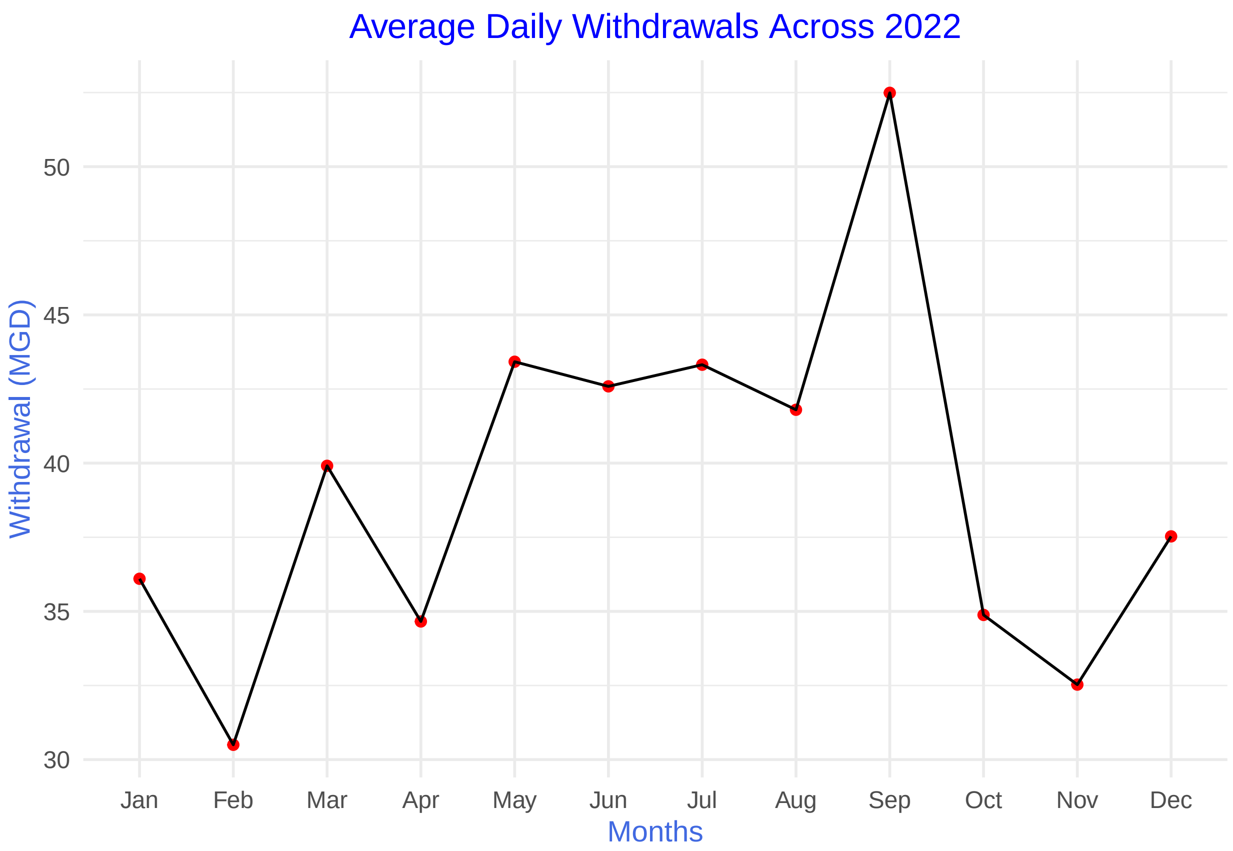
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   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 average daily withdrawals across the months for 2022

```
#4
months <- c(01, 05, 09, 02, 06, 10, 03, 07, 11, 04, 08, 12)
monthly.withdrawal.df <- data.frame(water.system.name,
                                    PWSID,
                                    ownership,
                                    months,
                                    max.withdrawals.mgd) %>%
  rename(withdrawal = max.withdrawals.mgd) %>%
  mutate(withdrawal = as.numeric(withdrawal),
         months = lubridate::month(months))

#5
axis.label <- as.vector(sort(lubridate::month(months, label = TRUE)))

ggplot(data = monthly.withdrawal.df,
       mapping = aes(x = months,
                     y = withdrawal)) +
  geom_point(colour = 'red') +
  geom_line(colour = 'black') +
  scale_x_discrete(limits = axis.label) +
  labs(x = "Months",
       y = "Withdrawal (MGD)",
       title = "Average Daily Withdrawals Across 2022")
```

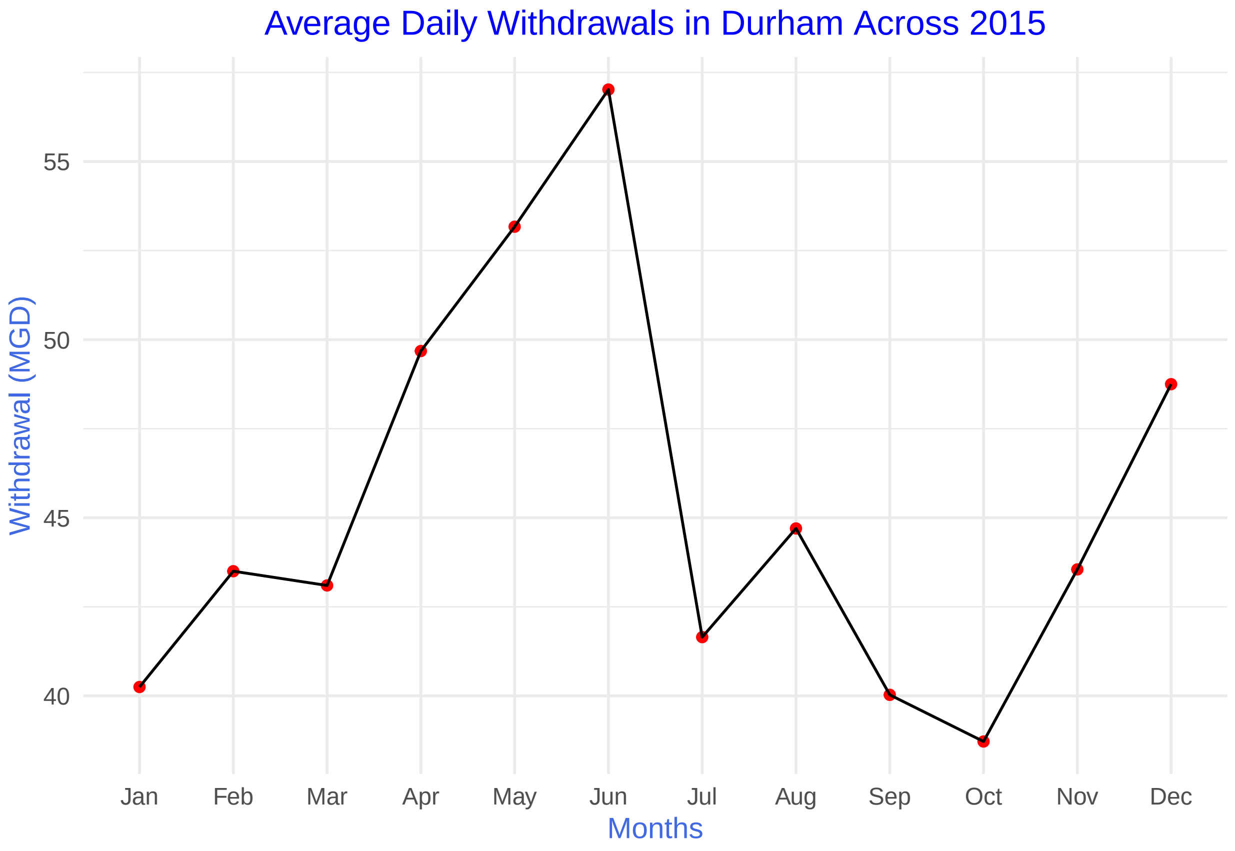

Average Daily Withdrawals Across 2022
Withdrawal (MGD)
50
45
40
35
30
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Months

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.

#https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2022
#url <- paste0("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=", url.pwsid, "&year=", url.year
#url

scraperboi <- function(url.pwsid, url.year) #This function takes the URL from ncdeq(), scrapes the webs
{
  url <- paste0("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=", url.pwsid, "&year=", url.yea

  water_supply <- read_html(url)

  #scraping the webpage for the data we need
  water.system <- water_supply %>%
    html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
    html_text()

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

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

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

  scaped.months <- c(01, 05, 09, 02, 06, 10, 03, 07, 11, 04, 08, 12)

  monthly.withdrawal <- data.frame(water.system,
                                   PWSID,
                                   ownership,
                                   url.year,
                                   scaped.months,
                                   max.withdrawals) %>%
    rename(withdrawal = max.withdrawals,
           months = scaped.months,
           year = url.year) %>%
    mutate(withdrawal = as.numeric(withdrawal),
           months = lubridate::month(months, label = FALSE))
  return(monthly.withdrawal)
}
```

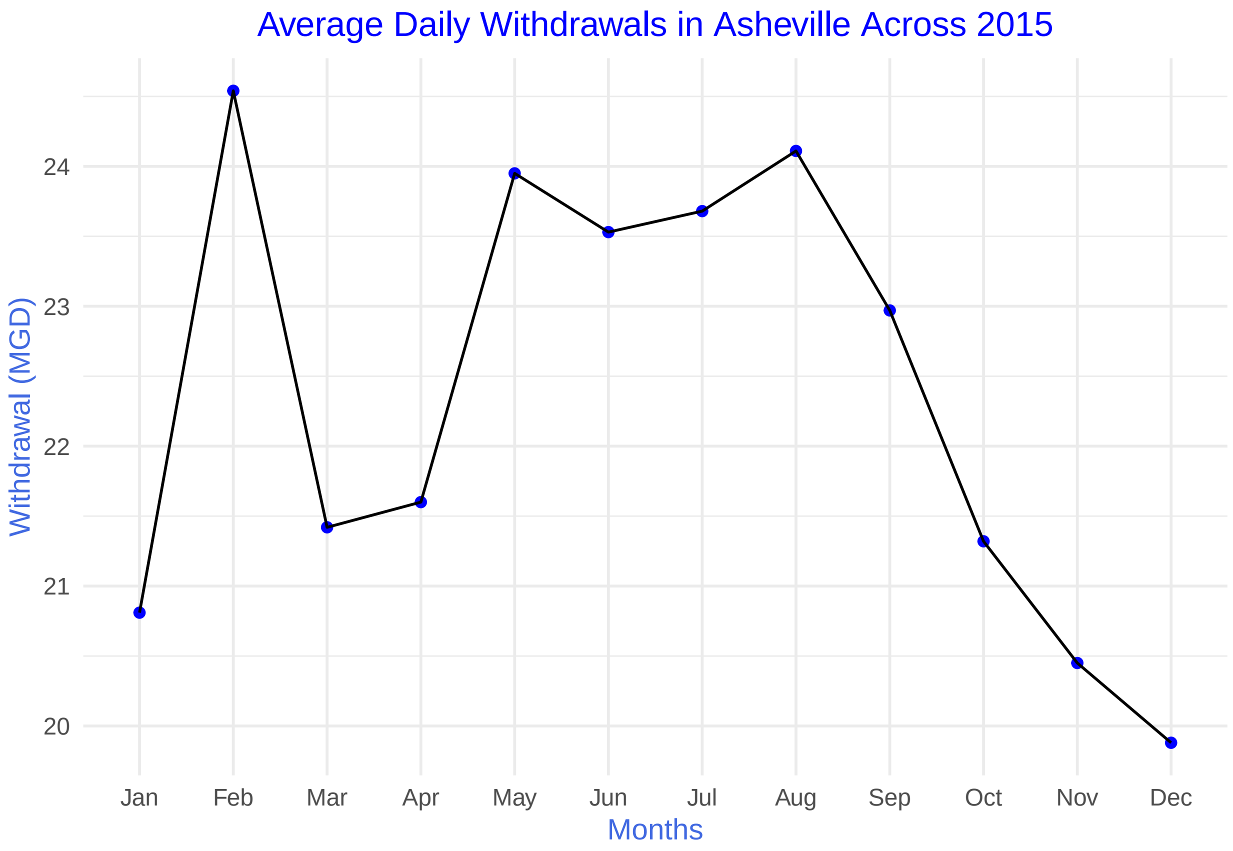
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID='03-32-010') for each month in 2015

```
#7
#these two variables serve as input arguments for ncdeq() which is called by scraperboi()
url.pwsid <- "03-32-010"
url.year <- "2015"

durham.monthly.withdrawal <- scraperboi(url.pwsid, url.year)

#plotting the results

ggplot(data = durham.monthly.withdrawal,
       mapping = aes(x = months,
                     y = withdrawal)) +
  geom_point(colour = 'red') +
  geom_line(colour = 'black') +
  scale_x_discrete(limits = axis.label) +
  labs(x = "Months",
       y = "Withdrawal (MGD)",
       title = "Average Daily Withdrawals in Durham Across 2015")
```



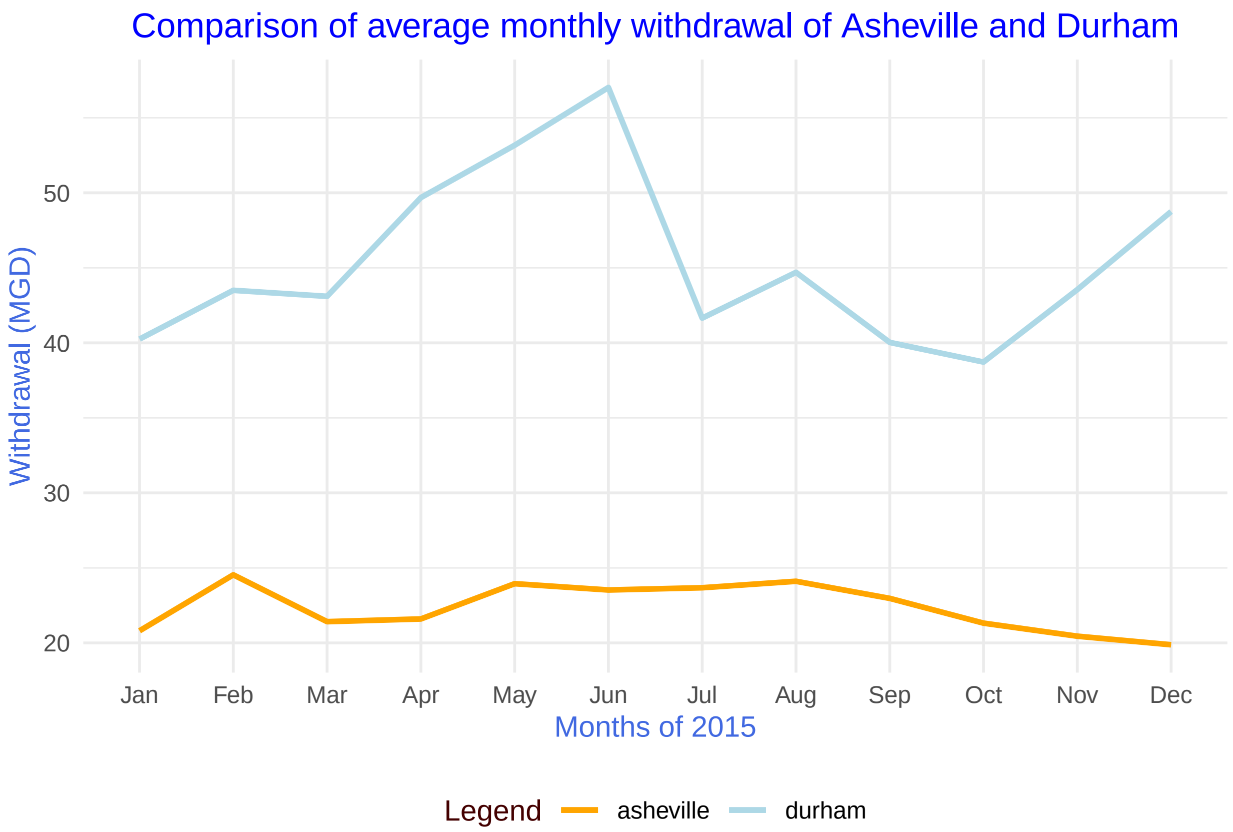
8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares Asheville's to Durham's water withdrawals.

```r
#8
asheville.monthly.withdrawal <- scraperboi('01-11-010', 2015)

#plotting the results

ggplot(data = asheville.monthly.withdrawal,
       mapping = aes(x = months,
                     y = withdrawal)) +
  geom_point(colour = 'blue') +
  geom_line(colour = 'black') +
  scale_x_discrete(limits = axis.label) +
  labs(x = "Months",
       y = "Withdrawal (MGD)",
       title = "Average Daily Withdrawals in Asheville Across 2015")
```



```r
legend.colours <- c("durham" = "lightblue", "asheville" = "orange")
ggplot() +
  geom_line(data = durham.monthly.withdrawal,
            mapping = aes(x = months,
                          y = withdrawal,
                          colour = "durham"),
            size = 1) +
  geom_line(data = asheville.monthly.withdrawal,
            mapping = aes(x = months,
                          y = withdrawal,
                           colour = 'asheville'),
            size = 1) +
  scale_x_discrete(limits = axis.label) +
  scale_color_manual(values = legend.colours) +
  labs(x = "Months of 2015",
       y = "Withdrawal (MGD)",
       title = "Comparison of average monthly withdrawal of Asheville and Durham",
       colour = "Legend")
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
```

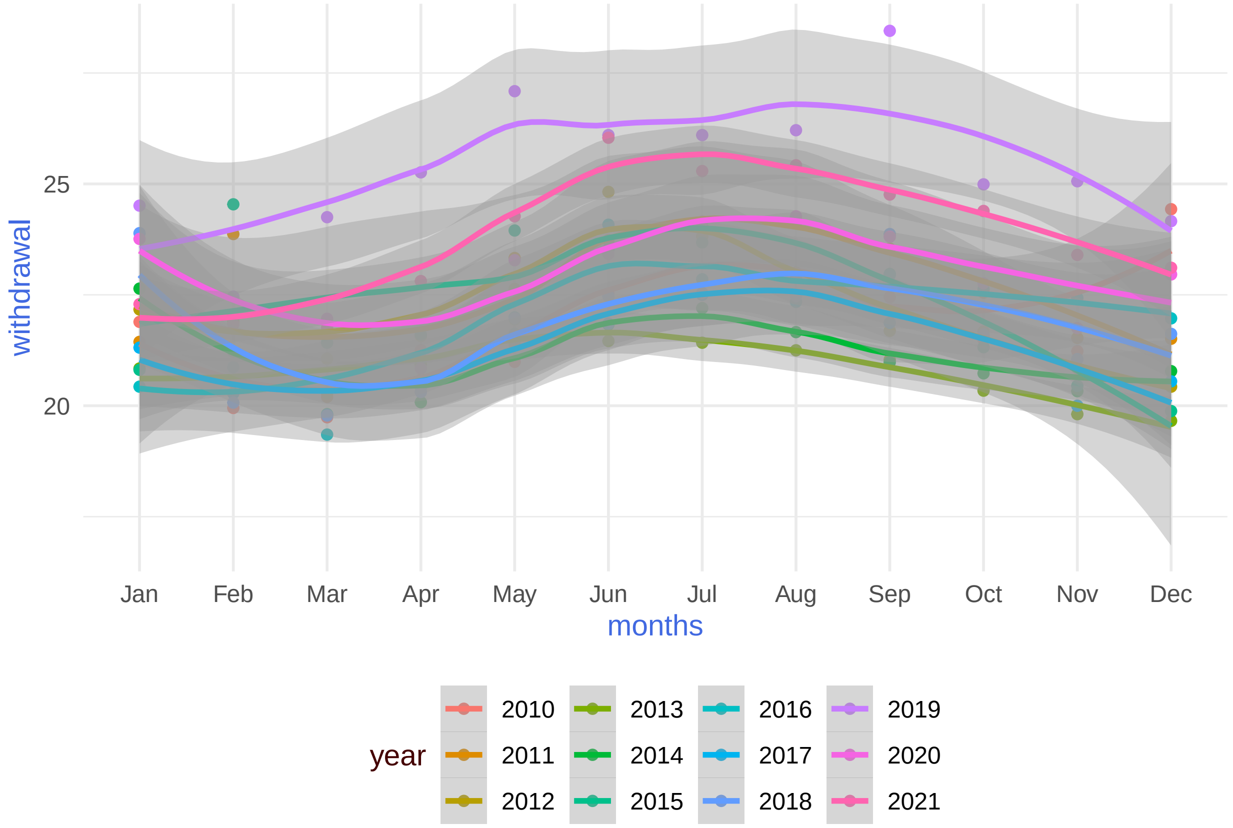


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2021.Add a smoothed line to the plot (method = 'loess').

   TIP: See Section 3.2 in the "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
years <- seq.int(from = 2010, to = 2021, by = 1)
location.pwsid <- rep("01-11-010", 12)

asheville.withdrawals.2010.2021 <- map2(location.pwsid,
                                        years,
                                        scraperboi) %>%
  bind_rows() %>%
  arrange(year, months) %>%
  mutate(year = as.factor(year))

ggplot(data = asheville.withdrawals.2010.2021,
       mapping = aes(x = months,
                     y = withdrawal,
                     colour = year)) +
  geom_point() +
  geom_smooth(method = "loess") +
  #geom_line() +
  scale_x_discrete(limits = axis.label)
```

```
## `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? We see that the water usage has a consistent seasonal component across the years in our sample size. Consumption is highest in July, decreases through autumn and winter to its lowest point in March.