Assignment 10: Data Scraping

Molly Bruce

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
getwd()
```

```
## [1] "C:/Users/mmb88/Desktop/Environmental_Data_Analytics_2021/Assignments"
```

```
# Load packages previously installed
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.3

## Warning: package 'ggplot2' was built under R version 4.0.3

## Warning: package 'tibble' was built under R version 4.0.3

## Warning: package 'tidyr' was built under R version 4.0.3

## Warning: package 'readr' was built under R version 4.0.3

## Warning: package 'purrr' was built under R version 4.0.3

## Warning: package 'dplyr' was built under R version 4.0.3

## Warning: package 'stringr' was built under R version 4.0.3
```

```
## Warning: package 'forcats' was built under R version 4.0.3
```

```
library(rvest)
```

```
## Warning: package 'rvest' was built under R version 4.0.3
## Warning: package 'xml2' was built under R version 4.0.3
```

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.0.3
```

```
# Set ggplot 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 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 URL to be scraped.

```
#2

# Indicate scraped URL
URL <- read_html('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2019')
URL
```

```
## {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 "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

# Scrape water system name and show it
water_system_name <- URL %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()
water_system_name
```

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

```r
# Scrape PSWID and show it
PSWID <- URL %>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
  html_text()
PSWID
```

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

```r
# Scrape ownership and show it
ownership <- URL %>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%
  html_text()
ownership
```

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

```r
# Scrape maximum monthly withdrawals and show it
mmw <- URL %>%
  html_nodes("th~ td+ td") %>%
  html_text()
mmw
```

```
##  [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"
```

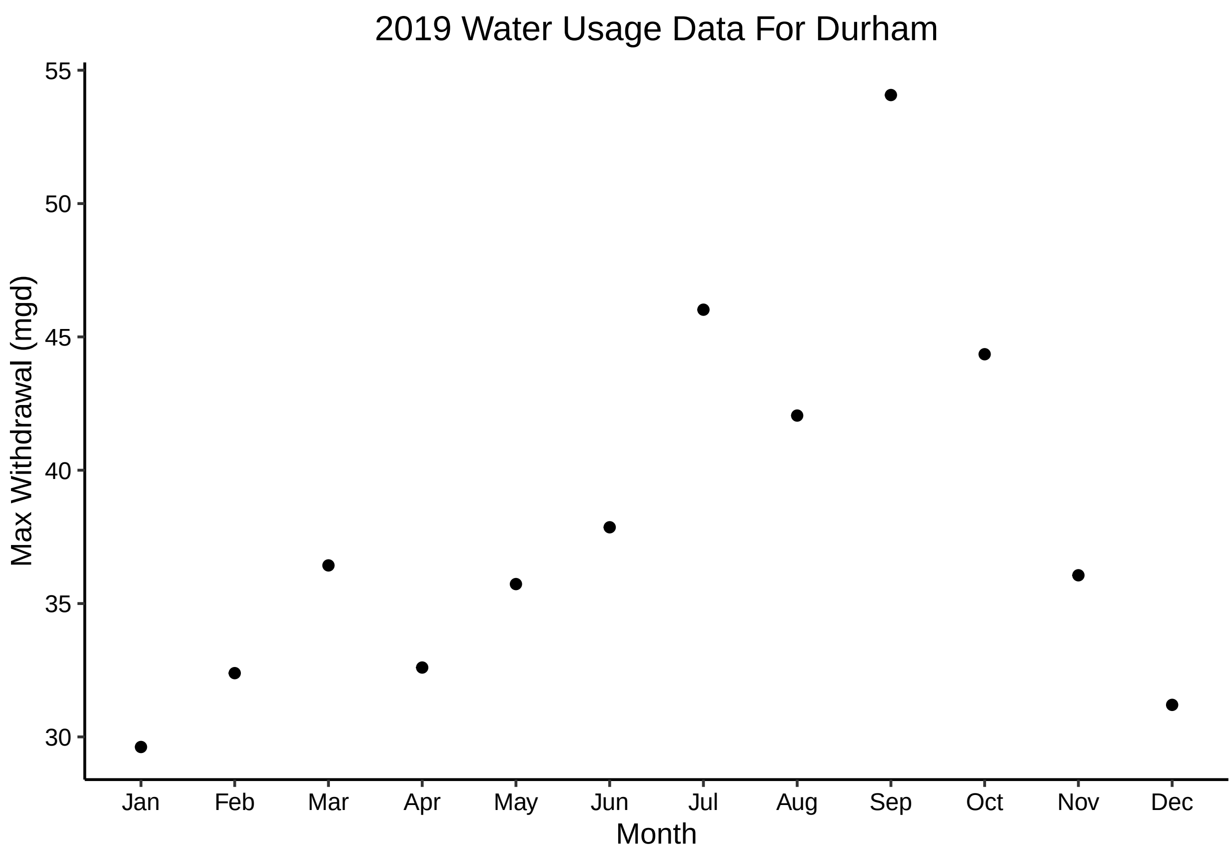
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

   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.

```r
#4

# Convert to a dataframe
dfURL <- data.frame("water_system_name" = water_system_name,
                    "PSWID" = PSWID,
                    "ownership" = ownership,
                    "maximum_withdrawl_MGD" = as.numeric(mmw),
                    # Add month column for data
                    "Month" = c('Jan','May','Sep','Feb','Jun','Oct',
                              'Mar','Jul','Nov','Apr','Aug','Dec'),
                    "MonthNum" = c('01','05','09','02','06','10',
                                 '03','07','11','04','08','12'),
                    # Add year column for data
                    "Year" = rep(2019,12),
                    # Add date column of month & year
                    Date = my(012019, 052019, 092019, 022019,
                             062019, 102019, 032019, 072019,
                             112019, 042019, 082019, 122019))

#5
```

```
# Order my Months in the 'csv' correctly (for plot purposes)
dfURL$Month <- factor(dfURL$Month, levels=c('Jan', 'Feb', 'Mar', 'Apr',
                                            'May', 'Jun', 'Jul', 'Aug',
                                            'Sep', 'Oct', 'Nov', 'Dec'))

# Plot by month
ggplot(dfURL,aes(x=Month,y=maximum_withdrawl_MGD)) +
  geom_point() +
  labs(title = paste("2019 Water Usage Data For",water_system_name),
       y="Max Withdrawal (mgd)",
       x="Month") +
  theme(plot.title = element_text(hjust=0.5))
```



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.

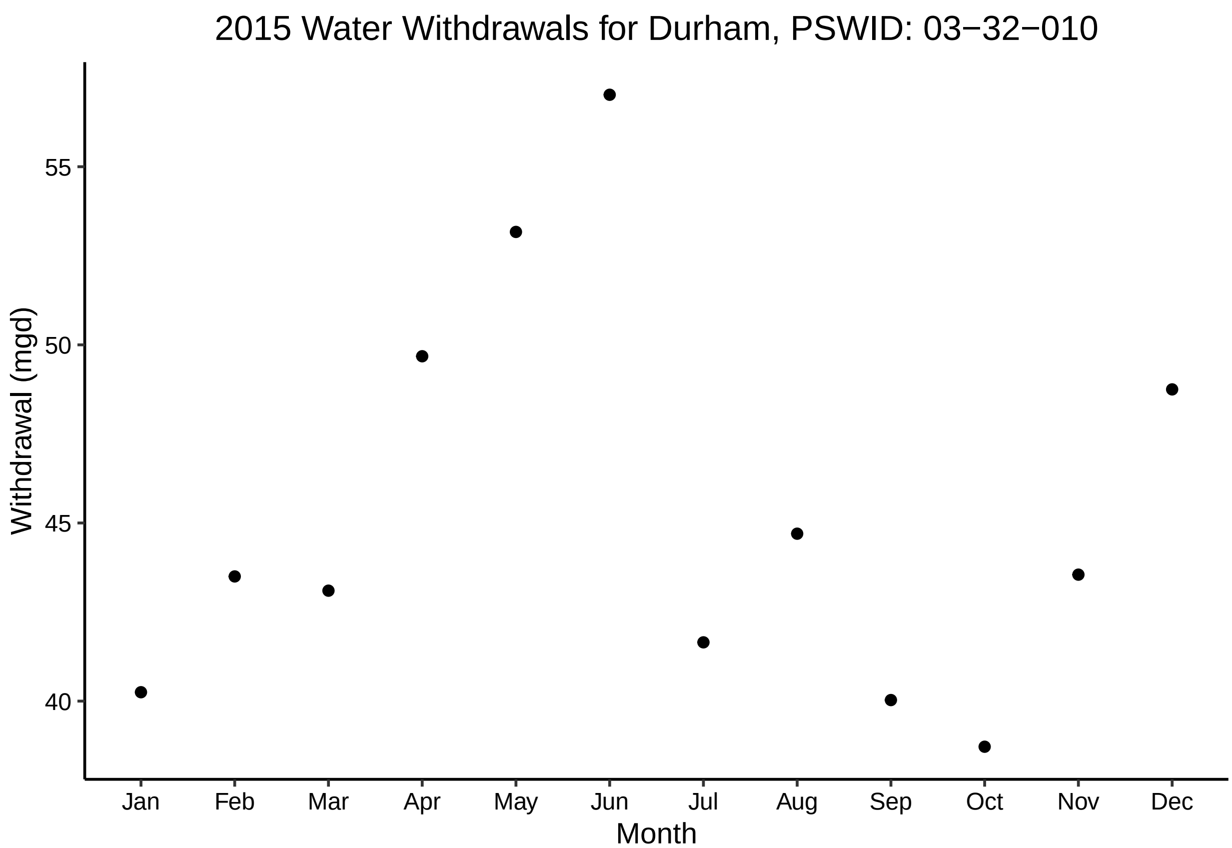
```
#6.

# Create scrape function
scrape <- function(year, PSWID){

  # Retrieve the website contents
  the_website <- read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',
  PSWID,'&year=',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)"
  mmw_tag <- "th~ td+ td"

  # Scrape the data items
  waterSystemName <- the_website %>% html_nodes(water_system_name_tag) %>% html_text()
  PSWIDValue <- the_website %>%   html_nodes(PSWID_tag) %>%  html_text()
  ownershipName <- the_website %>% html_nodes(ownership_tag) %>% html_text()
  mmwAmount <- the_website %>% html_nodes(mmw_tag) %>% html_text()

  # Convert to a dataframe
  dfURL2 <- data.frame("Month" = c('Jan','May','Sep','Feb','Jun','Oct',
                                  'Mar','Jul','Nov','Apr','Aug','Dec'),
                       "MonthNum" = c('01','05','09','02','06','10',
                                     '03','07','11','04','08','12'),
                       "Year" = rep(year,12),
                       "YearChar" = as.character(rep(year, 12))) %>%
    mutate("water_system_name" = !!waterSystemName,
           "PSWID" = !!PSWIDValue,
           "ownership" = !!ownershipName,
           "maximum_withdrawl_MGD" = as.numeric(!!mmwAmount),
           "Date" = my(paste0(MonthNum,"-",year)))

  # Order my Months in the 'csv' correctly
  dfURL2$Month <- factor(dfURL2$Month, levels=c('Jan', 'Feb', 'Mar', 'Apr',
                                                'May', 'Jun', 'Jul', 'Aug',
                                                'Sep', 'Oct', 'Nov', 'Dec'))

  #Return the dataframe
  return(dfURL2)

}
```

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

```
#7

# Create more useful inputs for function
baseURL <- 'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid='
PSWID <- '03-32-010'
year <- 2015
scrapeURL <- paste0(baseURL,PSWID,'&year=',year)
URL <- read_html(scrapeURL)
water_system_name_tag <- "div+ table tr:nth-child(1) td:nth-child(2)"
water_system_name <- URL %>% html_nodes(water_system_name_tag) %>% html_text()

# Run the function based on inputs and view output
thedfD <- scrape(year, PSWID)
view(thedfD)

# Order my Months in the 'csv' correctly
thedfD$Month <- factor(thedfD$Month, levels=c('Jan', 'Feb', 'Mar', 'Apr',
```

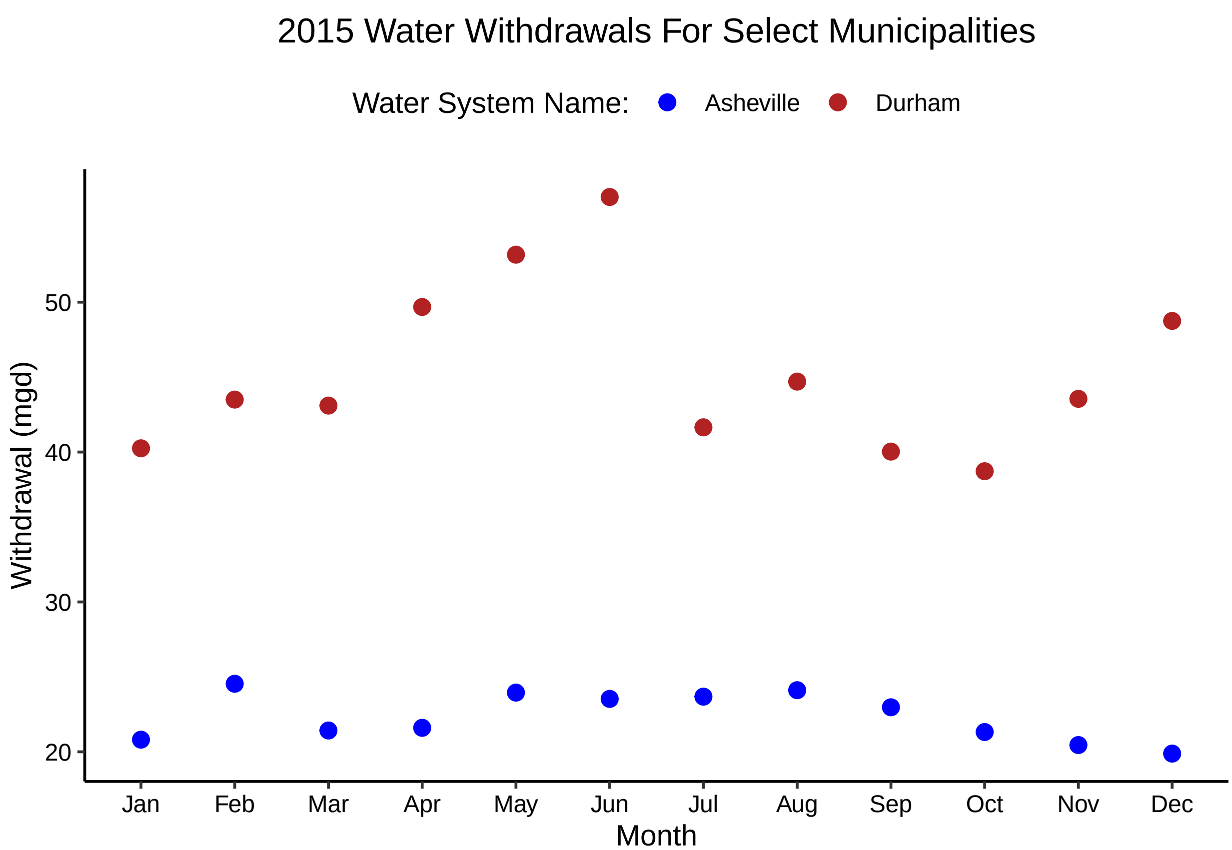
```
                                             'May', 'Jun', 'Jul', 'Aug',
                                             'Sep', 'Oct', 'Nov', 'Dec'))

# Plot
ggplot(thedfD,aes(x=Month,y=maximum_withdrawl_MGD)) +
  geom_point() +
  labs(title = paste0(year, " Water Withdrawals for ", water_system_name, ", PSWID: ", PSWID),
       y = "Withdrawal (mgd)",
       x = "Month") +
  theme(plot.title = element_text(hjust=0.5))
```



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

# Set inputs for Asheville
PSWID <- '01-11-010'
year <- 2015

# Run the function based on inputs and view output
thedfA <- scrape(year, PSWID)
view(thedfA)

# Combine location data to it
```

```r
thedfjoined <- rbind(thedfD, thedfA)
view(thedfjoined)

# Order my Months in the 'csv' correctly
thedfjoined$Month <- factor(thedfjoined$Month, levels=c('Jan', 'Feb', 'Mar', 'Apr',
                                                        'May', 'Jun', 'Jul', 'Aug',
                                                        'Sep', 'Oct', 'Nov', 'Dec'))

# Plot
ggplot(thedfjoined,
       aes(x=Month,y=maximum_withdrawl_MGD,color=water_system_name)) +
  geom_point(size = 2.5) +
  scale_color_manual(values = c("blue", "firebrick")) +
  labs(title = paste0(year, " Water Withdrawals For Select Municipalities"),
       y = "Withdrawal (mgd)",
       x = "Month",
       color = "Water System Name:") +
  theme(plot.title = element_text(hjust=0.5))
```

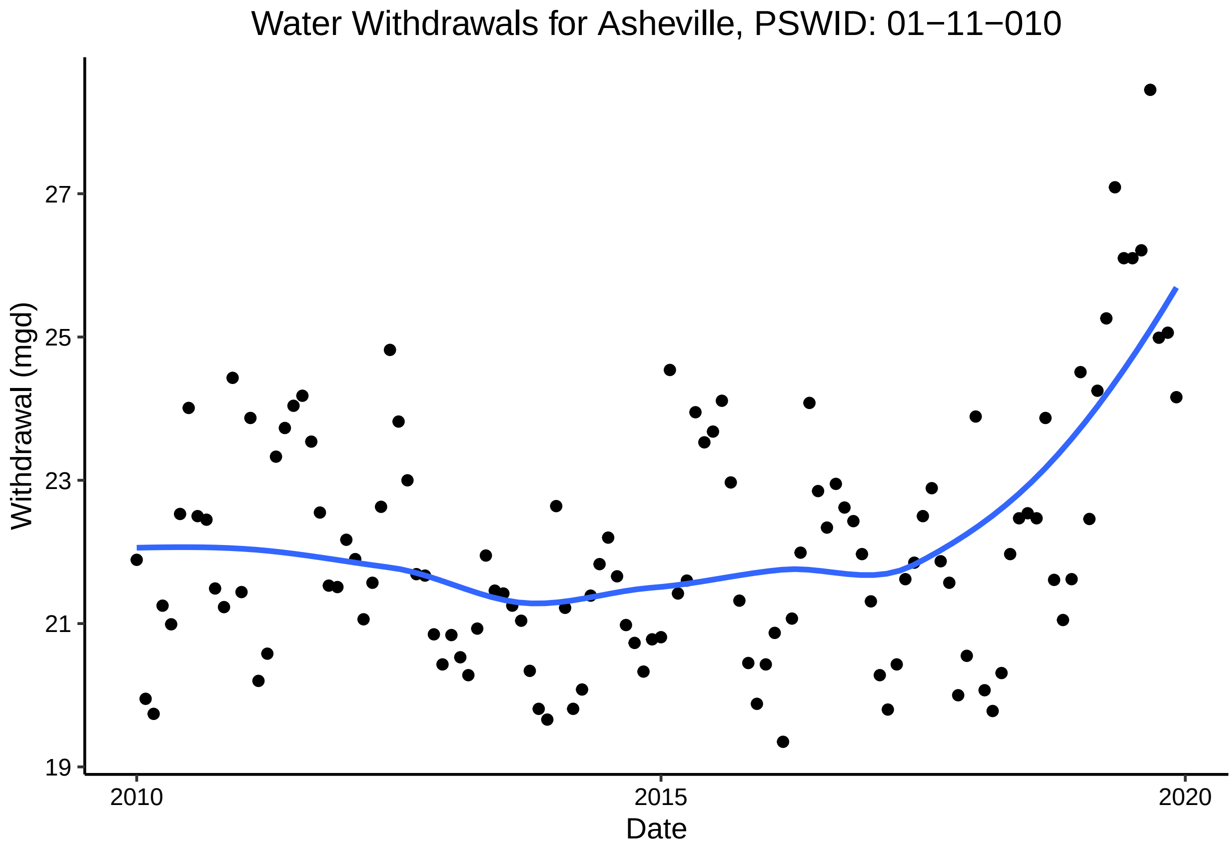


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

```r
#9

# Set inputs
baseURL <- 'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid='
```

```r
PSWID <- '01-11-010'
years <- rep(2010:2019)

# Add extra inputs for chart naming purposes
year <- 2010
scrapeURL <- paste0(baseURL,PSWID,'&year=',year)
URL <- read_html(scrapeURL)
water_system_name_tag <- "div+ table tr:nth-child(1) td:nth-child(2)"
water_system_name <- URL %>% html_nodes(water_system_name_tag) %>% html_text()

#Use lapply to apply the scrape function
thedfs <- lapply(X = years,
                 FUN = scrape,
                 PSWID=PSWID)

#Conflate the returned dataframes into a single dataframe
thedfAmore <- bind_rows(thedfs)
view(thedfAmore)

# Order my Months in the 'csv' correctly
thedfAmore$Month <- factor(thedfAmore$Month, levels=c('Jan', 'Feb', 'Mar', 'Apr',
                                            'May', 'Jun', 'Jul', 'Aug',
                                            'Sep', 'Oct', 'Nov', 'Dec'))

# Plot ASheville water useage from 2010 to 2019
ggplot(thedfAmore,
       aes(x=Date,y=maximum_withdrawl_MGD)) +
  geom_point() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste0("Water Withdrawals for ", water_system_name, ", PSWID: ", PSWID),
       y = "Withdrawal (mgd)",
       x = "Date") +
  theme(plot.title = element_text(hjust=0.5))
```

```
## `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?

The line of best fit indicates that, particularly over the past ~3 years, Asheville's water consumption has been increasing. Typically, industrial water users have a larger impact on usage patterns. Because this time line accords with the Trump administration's tenure and a general promotion of coal, my guess is that this increase in water usage might have something to do with increased power generation from coal and the ensuing water demands.