Assignment 09: Data Scraping

Shivani Kuckreja

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_09_Data_Scraping.Rmd") prior to submission.

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] "/Users/shivanikuckreja/OneDrive - Wellesley College/Duke/Spring 2022 Classes/Environmental Data
```

```
library(tidyverse)
library(lubridate)

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

# Set theme
mytheme <- theme_classic() +
  theme(axis.text = element_text(color = "blue"),
        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 2021 to 2020 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=2020

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=2020')
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 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, with the first value being 36.0100.

```
#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 order. You can overcome this by creating a month column in the same order the data are scraped: Jan, May, Sept, Feb, etc...

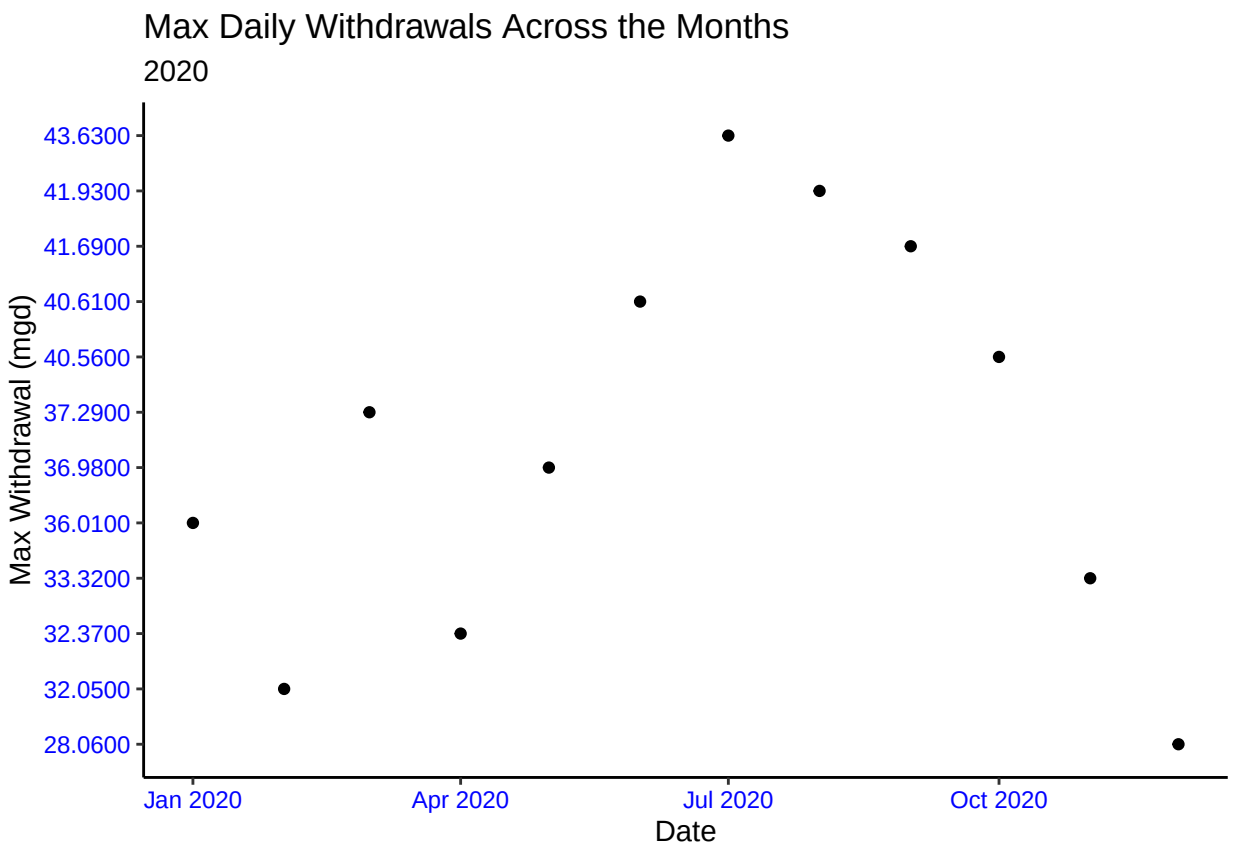
5. Plot the max daily withdrawals across the months for 2020

```
#4
Month <- webpage %>%
  html_nodes(".fancy-table:nth-child(31) tr+ tr th") %>% html_text()

Year <- 2020
```

```
the_df <- data.frame(
"Water System Name" = water.system.name,
"PWSID" = pwsid,
"Ownership" = ownership,
"Maximum Daily Use (MGD)" = max.withdrawals.mgd,
"Month" = Month,
"Year" = Year,
"Date" = my(paste(Month,"-",Year)
))


#5
#Plot
ggplot(the_df, aes(x=Date,y=max.withdrawals.mgd)) +
  geom_point() +
  labs(title = paste("Max Daily Withdrawals Across the Months"),
       subtitle = 2020,
      #scale_x_date(date_breaks="Month"), # code to show all month labels
      #on x-axis not working
       y="Max Withdrawal (mgd)",
       x="Date")
```




```
#geom_line not showing up which is why I am using geom_point
```

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 scraped**.

```
#6.
scrape.it <- function(Year, PWSID){
  the_url <-
    read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',
    PWSID, '&year=',Year))
    print(the_url)


  water.system.name_revised <- the_url %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>% html_text()
  pwsid_revised <- the_url %>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)") %>% html_text()
  ownership_revised <- the_url %>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>% html_text()
  max.withdrawals.mgd_revised <- the_url %>% html_nodes("th~ td+ td") %>% html_text()


max.withdrawls.year.new <- Year

dataframe <- data.frame("WaterSystemName" = as.character(water.system.name_revised),
"PWSID" =  as.character(pwsid_revised),
"Ownership" =  as.character(ownership_revised),
"MaxWithdrawals" =  as.character(max.withdrawals.mgd_revised),
"Month" =  as.character(Month),
"Years" = max.withdrawls.year.new)

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

return(dataframe)
}
```

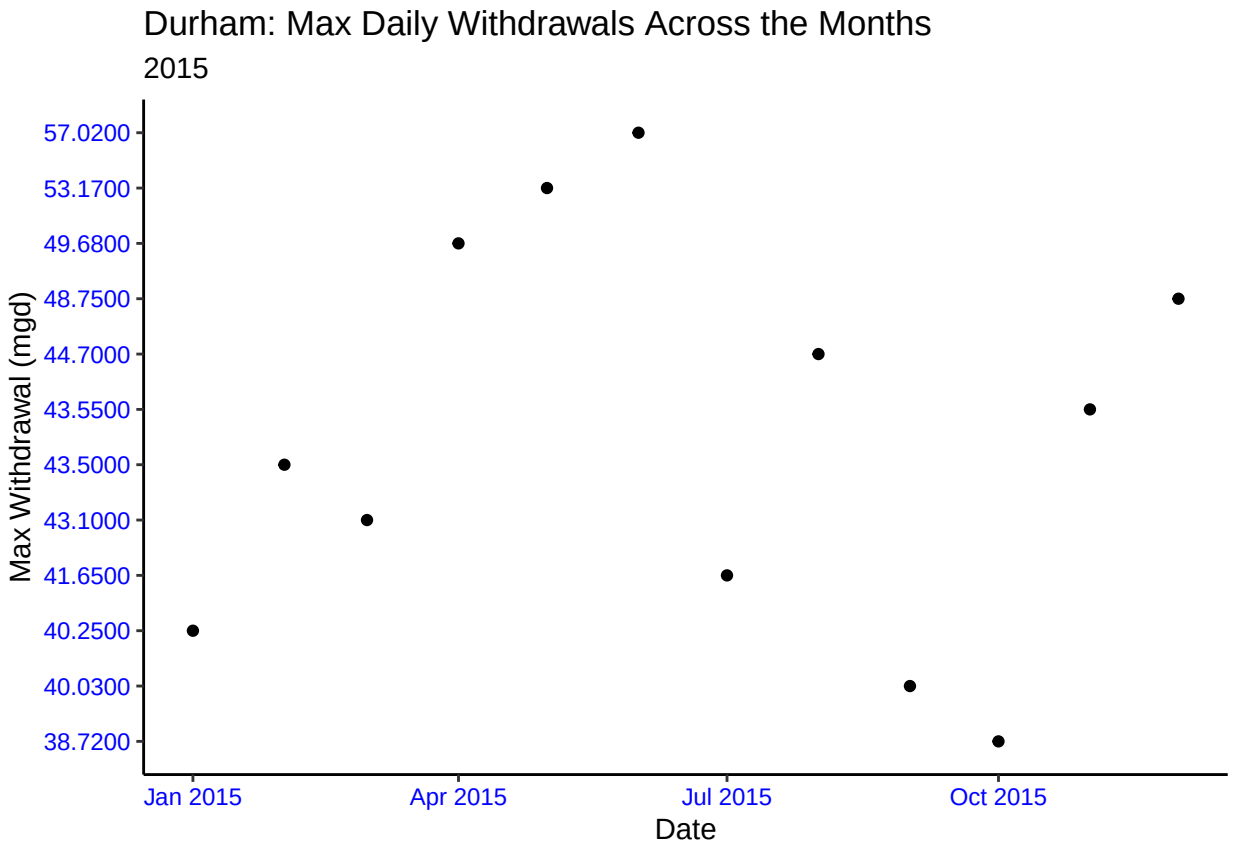
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID='03-32-010') for each month in 2015

```
#7
#Similar to 5 except use function

withdrawals.2015 <- scrape.it (2015, '03-32-010')
```

```
## {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= ...
```

```
ggplot(withdrawals.2015, aes(x=Date,y=MaxWithdrawals)) +
  geom_point() +
  labs(title = paste("Durham: Max Daily Withdrawals Across the Months"),
       subtitle = 2015,
      #scale_x_date(date_breaks="Month"), # code to show all month labels
      #on x-axis not working
       y="Max Withdrawal (mgd)",
       x="Date")
```



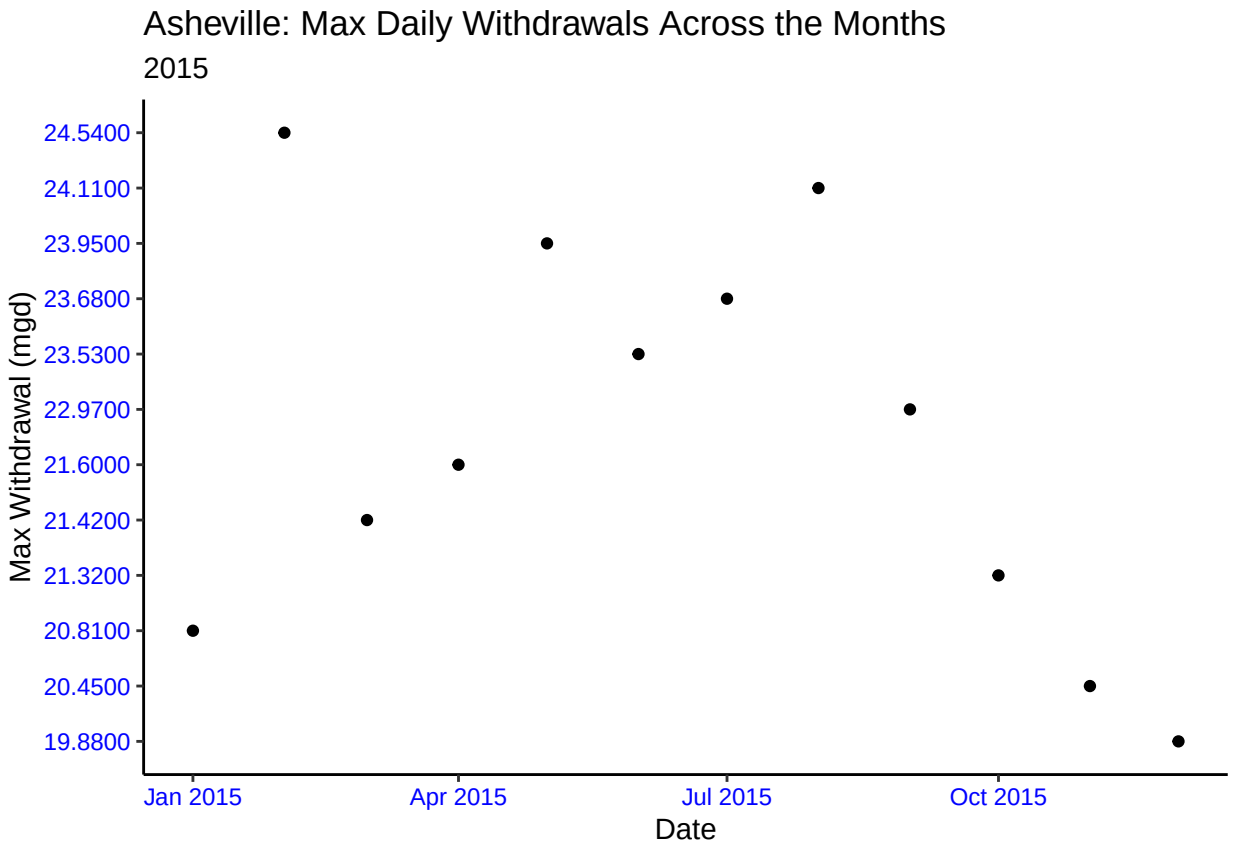

```
#geom_line not showing up which is why I am using geom_point
```

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
Asheville.2015 <- scrape.it (2015, '01-11-010')
```

```
## {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= ...
```

```
ggplot(Asheville.2015, aes(x=Date,y=MaxWithdrawals)) +
  geom_point() +
  labs(title = paste("Asheville: Max Daily Withdrawals Across the Months"),
       subtitle = 2015,
       y="Max Withdrawal (mgd)",
       x="Date")
```

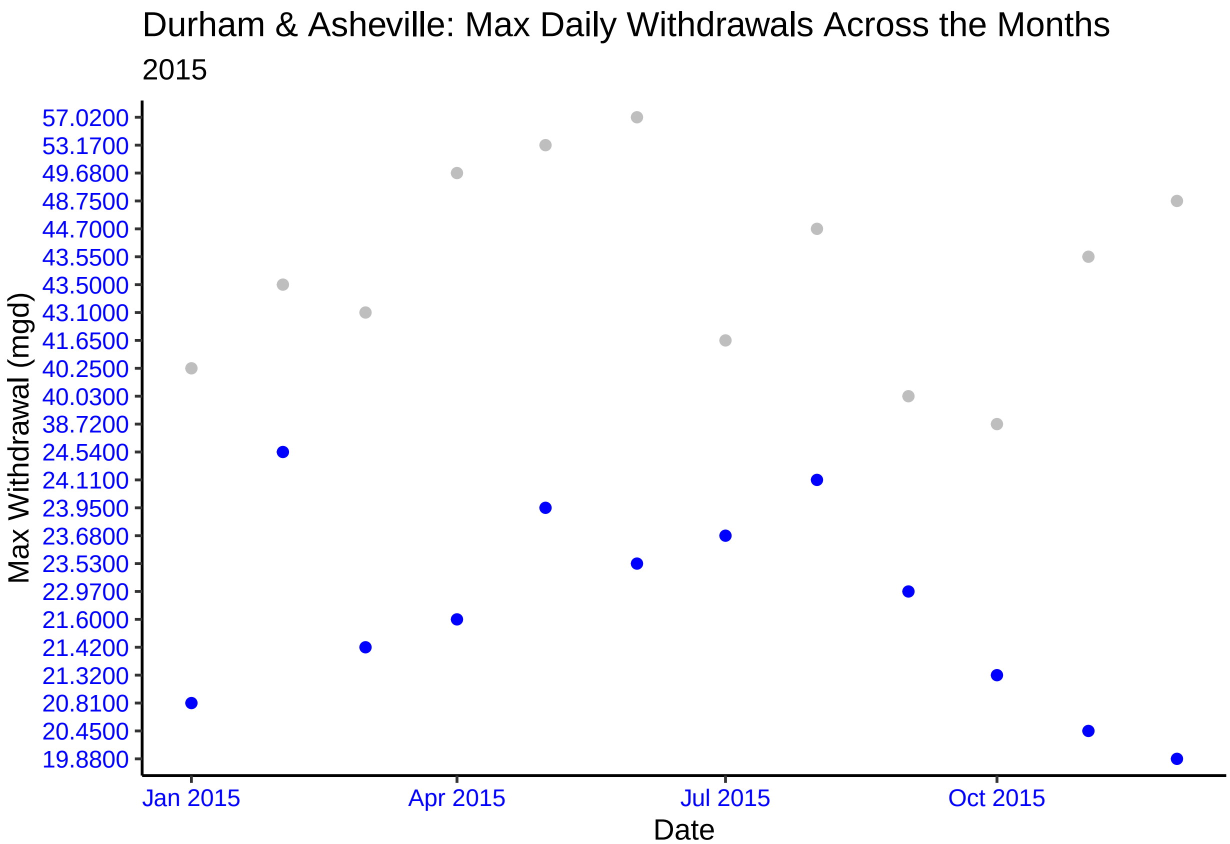



```
#geom_line not showing up which is why I am using geom_point

#Combine graphs

Newgraph <- rbind(Asheville.2015,withdrawals.2015)


ggplot(Newgraph,aes(x=Date,y=MaxWithdrawals)) +
  geom_point(data = Asheville.2015, color ='blue') +
   geom_point(data = withdrawals.2015, color ='grey') +
  labs(title = paste("Durham & Asheville: Max Daily Withdrawals Across the Months"),
       subtitle = 2015,
       y="Max 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 2020.Add a smoothed line to the plot.

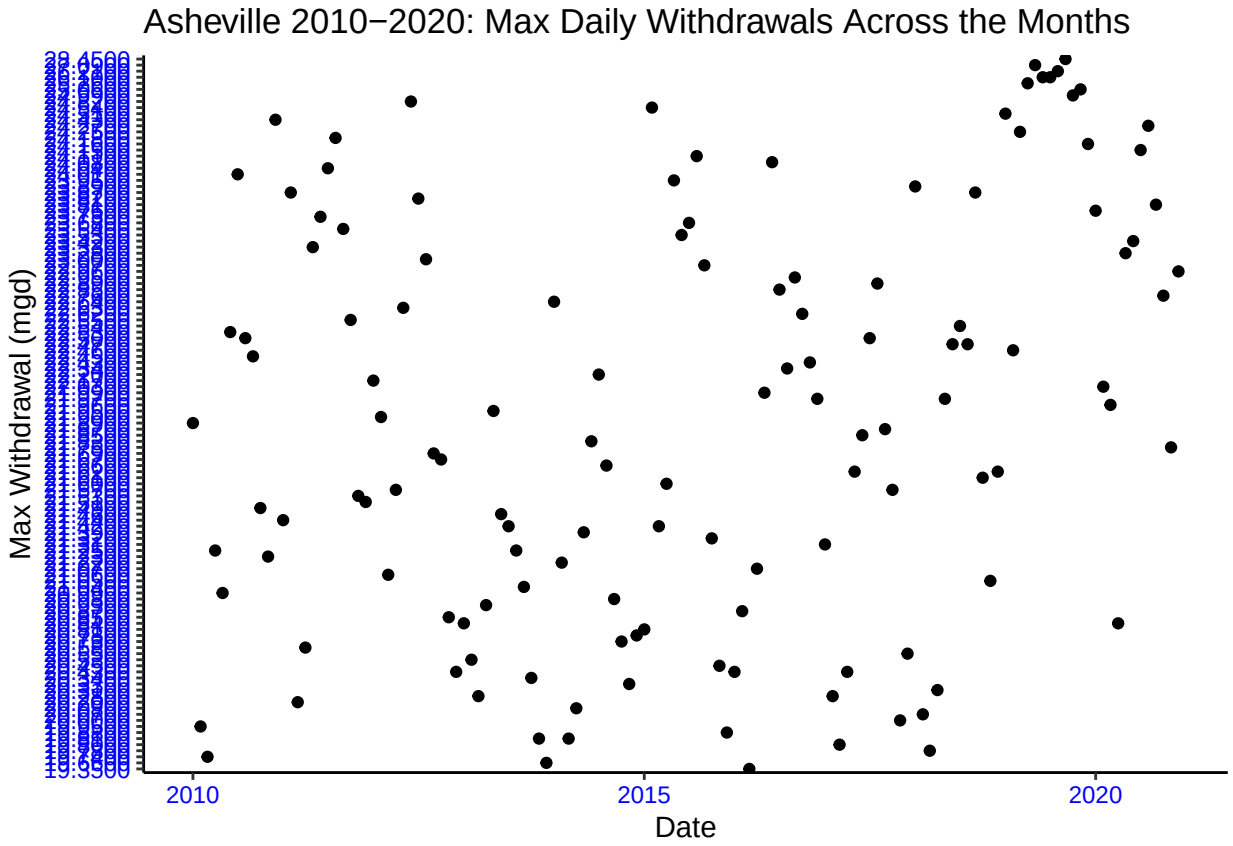
```
#9
Year = seq(2010, 2020)
pwsid = '01-11-010'

Ashville2010to2020 <- Year %>%
  map(scrape.it, PWSID = '01-11-010') %>%
  bind_rows()
```

```
## {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= ...
## {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= ...
## {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= ...
## {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= ...
## {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= ...
## {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= ...
## {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= ...
## {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= ...
## {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= ...
## {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= ...
## {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= ...
```

```
ggplot(Ashville2010to2020,aes(x=Date,y=MaxWithdrawals)) +
  geom_point() +
  labs(title = paste("Asheville 2010-2020: Max Daily Withdrawals Across the Months"),
       y="Max Withdrawal (mgd)",
       x="Date")
```



Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time? No