Assignment 09: Data Scraping

Matthew Vining

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

setwd("~/R/EDA-Fall2022")
getwd()
```

```
## [1] "/home/guest/R/EDA-Fall2022"
```

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

```
## {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".

```
#3

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

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

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

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

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

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

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

```
##  [1] "27.6400" "41.7900" "36.7200" "27.9700" "37.9500" "42.2400" "30.5400"
##  [8] "43.6200" "31.2800" "33.7600" "46.0800" "29.7800"
```

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.

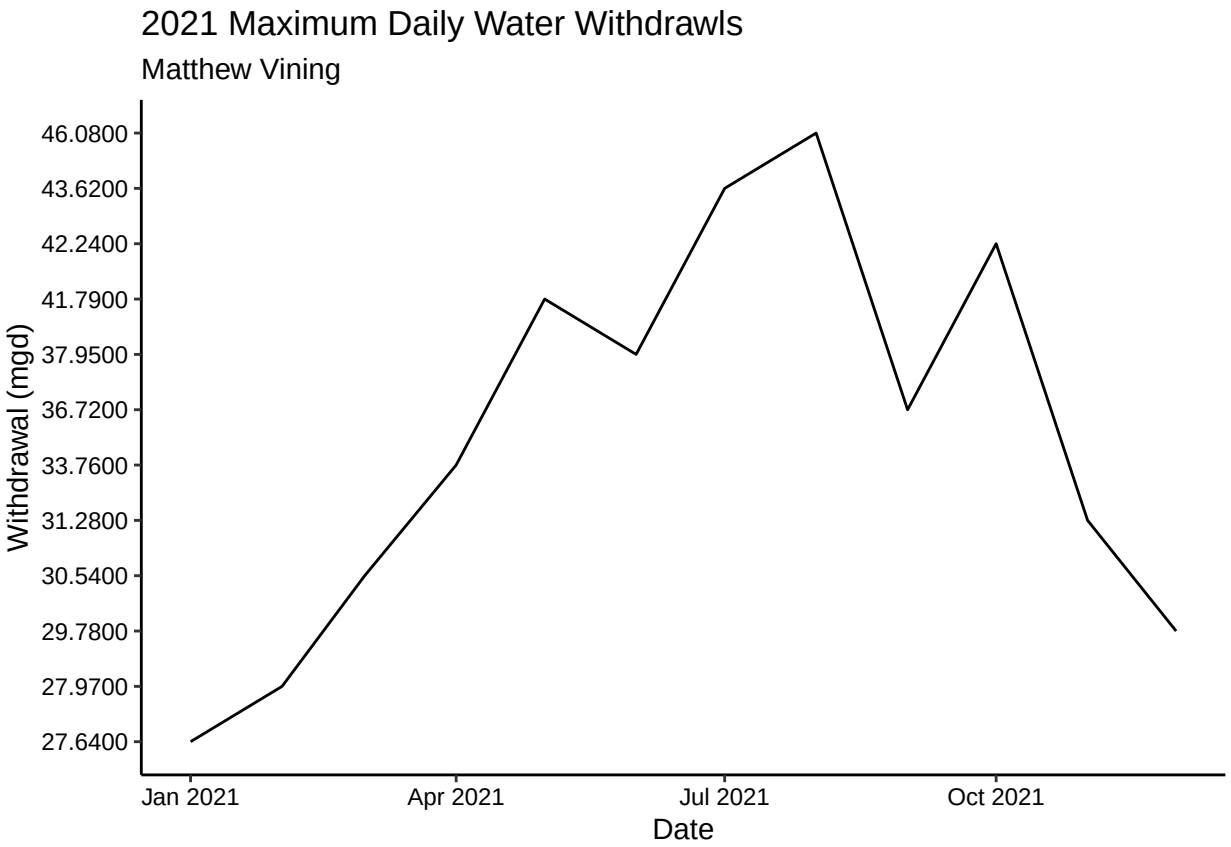
NOTE: It's likely you won't be able to scrape the monthly widthrawal data in chronological order. You can overcome this by creating a month column manually assigning values in the order the data are scraped: "Jan", "May", "Sept", "Feb", etc...

5. Create a line plot of the maximum daily withdrawals across the months for 2021

```
#4
df_withdrawals <- data.frame(water.system.name = water.system.name,
                             pswid = pswid,
                             ownership = ownership,
                             max.withdrawals.mgd = max.withdrawals.mgd,
                            "Month" = c("Jan", "May", "Sep", "Feb", "Jun",
                                        "Oct", "Mar", "Jul", "Nov", "Apr", "Aug", "Dec"),
                            "Year" = rep(2021,1))

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


#5
ggplot(df_withdrawals,aes(x=Date,y=max.withdrawals.mgd)) +
  geom_line(aes(group=1)) +
  labs(title = paste("2021 Maximum Daily Water Withdrawls"),
       subtitle = "Matthew Vining",
       y="Withdrawal (mgd)",
       x="Date")
```



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.
the_base_url <- 'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid='
pswid <- '03-32-010'
Year <- 2021
DEQwebpage <- paste0(the_base_url, pswid, '&year=', Year)
print(DEQwebpage)
```

```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2021"
```

```
scrape.it <- function(Year, pswid){

  #Retrieve the website contents
  DEQwebpage <- read_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',
                                 pswid, '&year=', Year))

  #Set the elements and scrape

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

  pswid <- DEQwebpage %>%
```

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

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

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

  #Convert to a dataframe
  df_withdrawals <- data.frame(water.system.name = water.system.name,
                               pswid = pswid,
                               ownership = ownership,
                               max.withdrawals.mgd = max.withdrawals.mgd,
                              "Month" = c("Jan", "May", "Sep", "Feb", "Jun",
                                          "Oct", "Mar", "Jul", "Nov", "Apr", "Aug", "Dec"),
                              "Year" = rep(Year,1))

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

  #Pause for a moment - scraping etiquette
  #Sys.sleep(1) #uncomment this if you are doing bulk scraping!

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

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

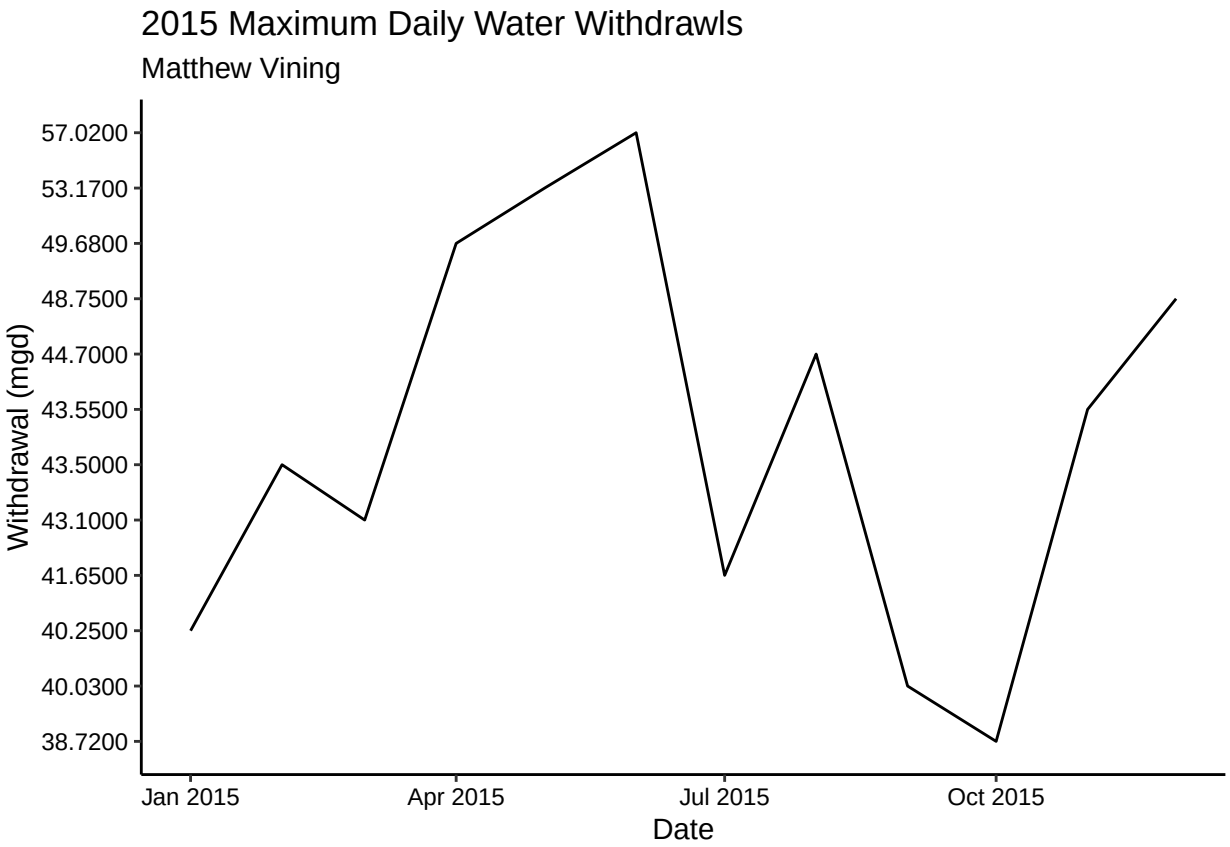
```
#7

extract.df <- scrape.it(Year=2015,pswid='03-32-010')
extract.df
```

```
##    water.system.name     pswid    ownership max.withdrawals.mgd Month Year
## 1             Durham 03-32-010 Municipality             40.2500   Jan 2015
## 2             Durham 03-32-010 Municipality             43.5000   Feb 2015
## 3             Durham 03-32-010 Municipality             43.1000   Mar 2015
## 4             Durham 03-32-010 Municipality             49.6800   Apr 2015
## 5             Durham 03-32-010 Municipality             53.1700   May 2015
## 6             Durham 03-32-010 Municipality             57.0200   Jun 2015
## 7             Durham 03-32-010 Municipality             41.6500   Jul 2015
## 8             Durham 03-32-010 Municipality             44.7000   Aug 2015
## 9             Durham 03-32-010 Municipality             40.0300   Sep 2015
## 10            Durham 03-32-010 Municipality             38.7200   Oct 2015
## 11            Durham 03-32-010 Municipality             43.5500   Nov 2015
## 12            Durham 03-32-010 Municipality             48.7500   Dec 2015
```

```
##          Date
## 1  2015-01-01
## 2  2015-02-01
## 3  2015-03-01
## 4  2015-04-01
## 5  2015-05-01
## 6  2015-06-01
## 7  2015-07-01
## 8  2015-08-01
## 9  2015-09-01
## 10 2015-10-01
## 11 2015-11-01
## 12 2015-12-01
```

```
ggplot(extract.df,aes(x=Date,y=max.withdrawals.mgd)) +
  geom_line(aes(group=1)) +
  labs(title = paste("2015 Maximum Daily Water Withdrawls"),
       subtitle = "Matthew Vining",
       y="Withdrawal (mgd)",
       x="Date")
```



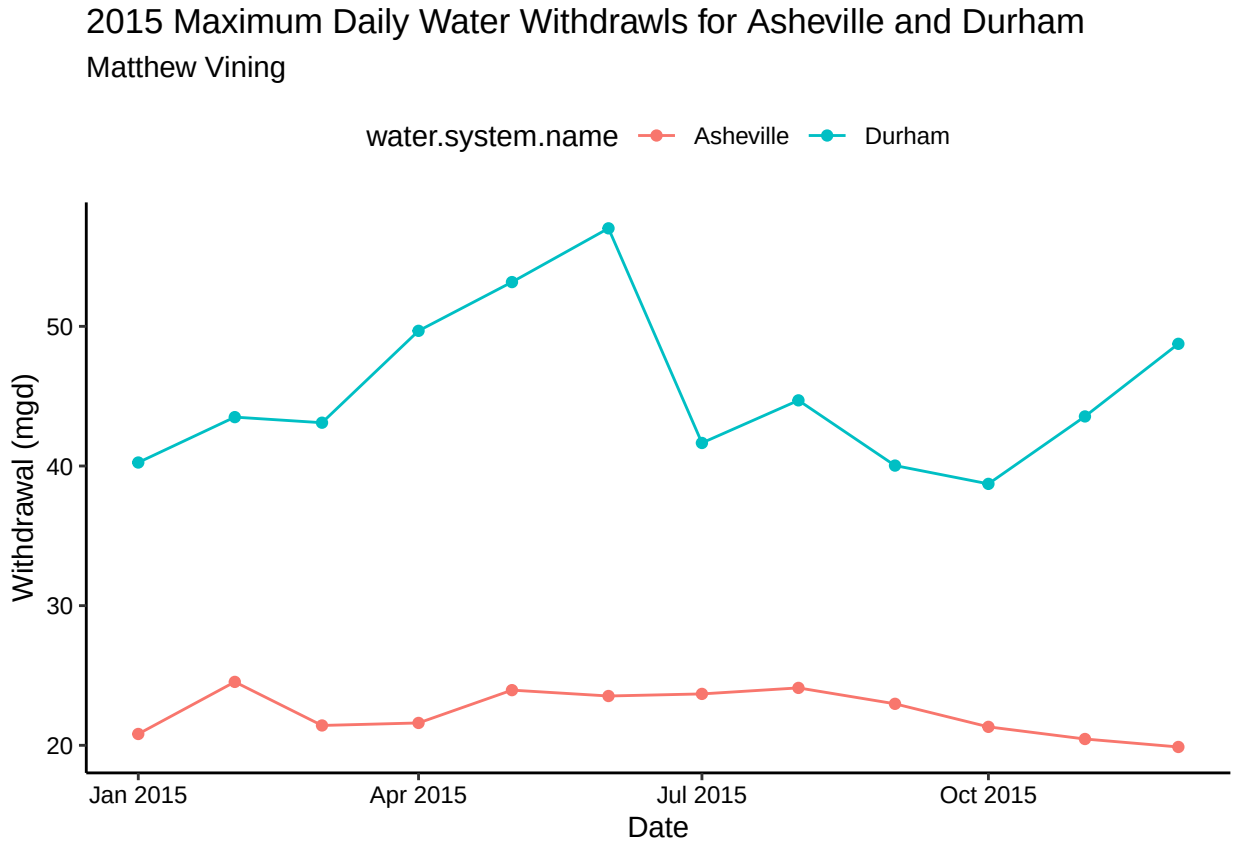
8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares Asheville's to Durham's water withdrawals.

```
#8
extract.df2 <- scrape.it(Year=2015,pswid='01-11-010')
extract.df2
```

```
##    water.system.name     pswid    ownership max.withdrawals.mgd Month Year
## 1          Asheville 01-11-010 Municipality             20.8100   Jan 2015
## 2          Asheville 01-11-010 Municipality             24.5400   Feb 2015
## 3          Asheville 01-11-010 Municipality             21.4200   Mar 2015
## 4          Asheville 01-11-010 Municipality             21.6000   Apr 2015
## 5          Asheville 01-11-010 Municipality             23.9500   May 2015
## 6          Asheville 01-11-010 Municipality             23.5300   Jun 2015
## 7          Asheville 01-11-010 Municipality             23.6800   Jul 2015
## 8          Asheville 01-11-010 Municipality             24.1100   Aug 2015
## 9          Asheville 01-11-010 Municipality             22.9700   Sep 2015
## 10         Asheville 01-11-010 Municipality             21.3200   Oct 2015
## 11         Asheville 01-11-010 Municipality             20.4500   Nov 2015
## 12         Asheville 01-11-010 Municipality             19.8800   Dec 2015
##          Date
## 1  2015-01-01
## 2  2015-02-01
## 3  2015-03-01
## 4  2015-04-01
## 5  2015-05-01
## 6  2015-06-01
## 7  2015-07-01
## 8  2015-08-01
## 9  2015-09-01
## 10 2015-10-01
## 11 2015-11-01
## 12 2015-12-01
```

```
total_extract <- rbind(extract.df, extract.df2)
total_extract$max.withdrawals.mgd <- as.numeric(total_extract$max.withdrawals.mgd)

ggplot(total_extract,aes(x=Date)) +
  geom_line(aes(y=max.withdrawals.mgd, color=water.system.name)) +
  geom_point(aes(y=max.withdrawals.mgd, color=water.system.name)) +
  labs(title = paste("2015 Maximum Daily Water Withdrawls for Asheville and Durham"),
       subtitle = "Matthew Vining",
       y="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 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.

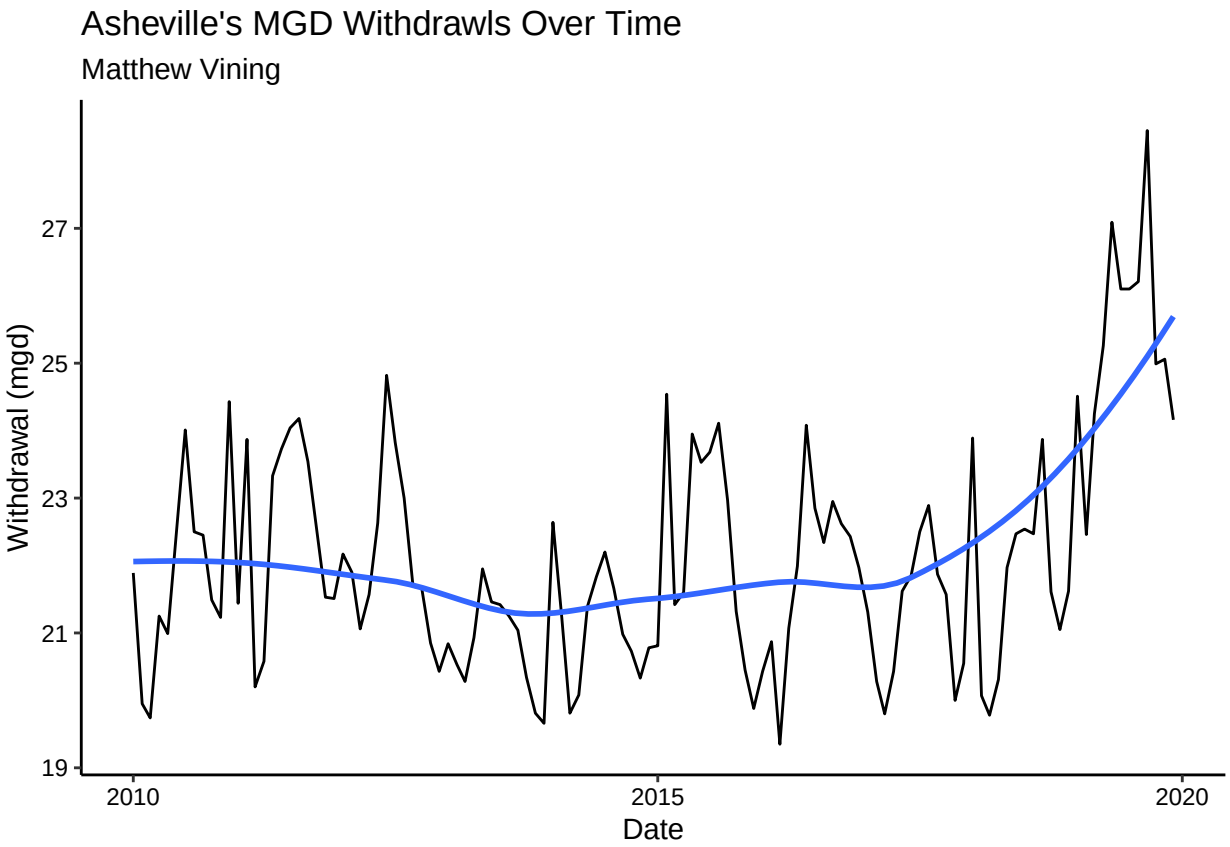
```
#9
Year = rep(2010:2019)
pswid = '01-11-010'

map_9 <- map(Year,scrape.it,pswid=pswid)
single_map_9 <- bind_rows(map_9)

single_map_9$max.withdrawals.mgd <- as.numeric(single_map_9$max.withdrawals.mgd)

ggplot(single_map_9,aes(x=Date,y=max.withdrawals.mgd)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("Asheville's MGD Withdrawls Over Time"),
       subtitle = "Matthew Vining",
       y="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, throughout time, water usage has had an increasing trend upwards, indicating greater use over time.