Assignment 10: Data Scraping

Andreana Chou

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

```r
#1

library(tidyverse)
library(lubridate)
library(rvest)
library(here)

here()
```

```
## [1] "C:/Users/andre/Documents/R Studio Files/EDE_Fall2023"
```

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

#used read_html() and copied/pasted the url
durham.lwsp <- 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)“.

```
#3

#scrape water system name
Durham <- durham.lwsp %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()

#scrape PWSID
durham.pwsid <- durham.lwsp %>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
  html_text()

#scrape ownership
durham.ownership <- durham.lwsp %>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%
  html_text()

#scrape maximum day use per month
durham.mgd <- durham.lwsp %>%
  html_nodes("th~ td+ td") %>%
  html_text()
```

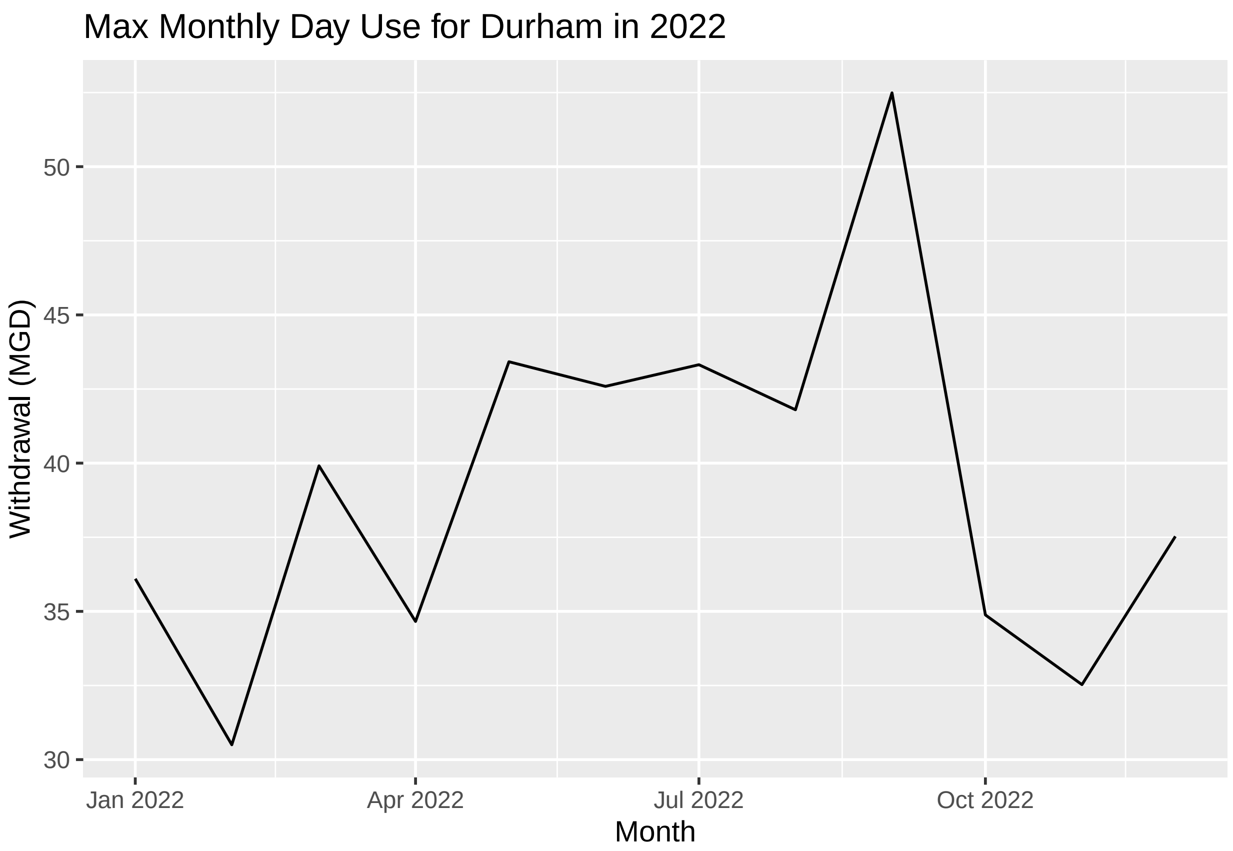
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

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 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

#used data.frame(), set "Month" as a string with the correct order of months,
#set maximum withdrawals as a numeric data type
durham.withdrawals <- data.frame("Month" = c("Jan", "May", "Sep", "Feb", "Jun",
                                             "Oct", "Mar", "Jul", "Nov",
                                             "Apr", "Aug", "Dec"),
                                 "Year" = rep(2022, 12),
                                 "Max_Withdrawals" = as.numeric(durham.mgd))

#mutated columns in data frame to set static attributes as variables, created
#Date column
durham.withdrawals <- durham.withdrawals %>%
  mutate(City = !!Durham,
         PWSID = !!durham.pwsid,
         Ownership = !!durham.ownership,
         Date = my(paste(Month,"-",Year)))

#5

#used ggplot() and geom_line() to create line plot of scraped data
durham.2022.plot <- durham.withdrawals %>%
  ggplot(aes(x=Date, y=Max_Withdrawals)) +
  geom_line() +
  labs(title=paste("Max Monthly Day Use for",Durham,"in 2022"),
       x = "Month",
       y = "Withdrawal (MGD)")
durham.2022.plot
```

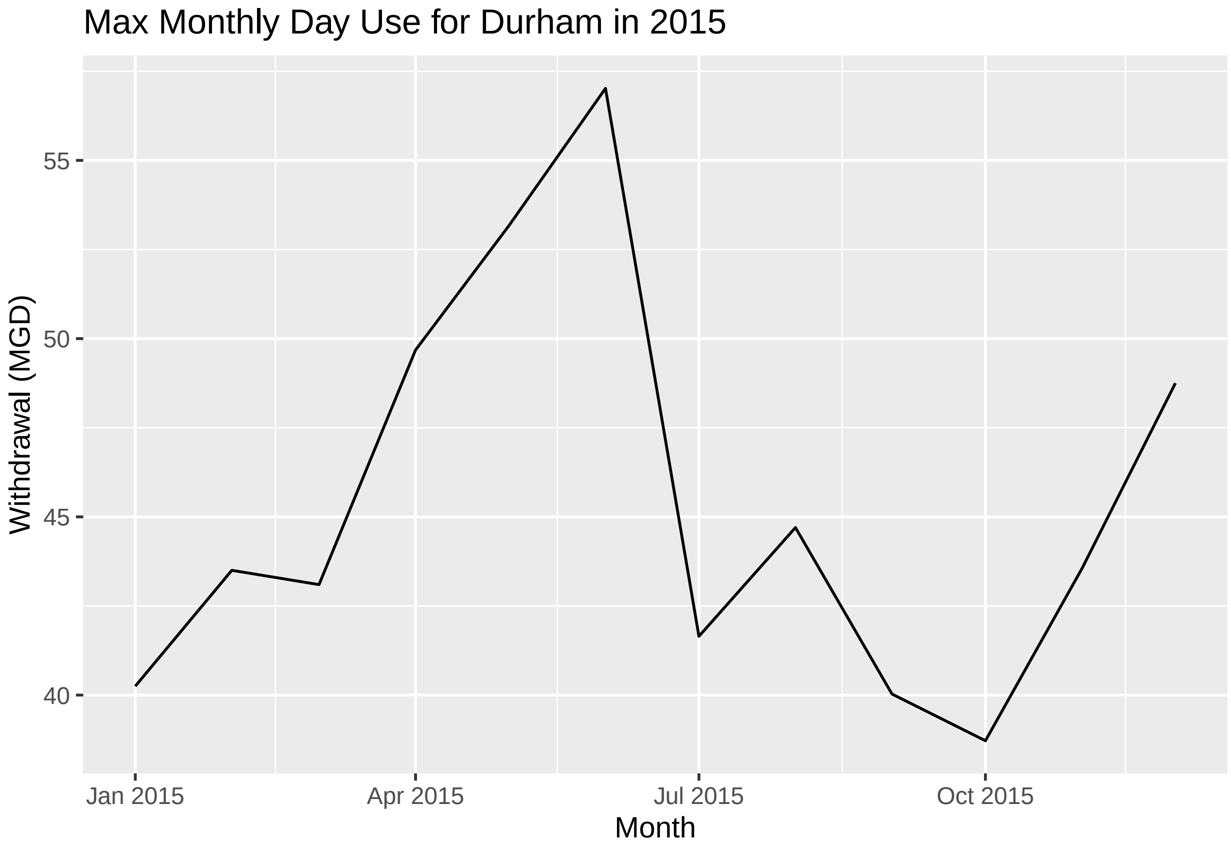


6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

```r
#6.

#used function with two variables: pwsid and year
scrape.page <- function(url_pwsid, url_year) {
  #scraping web address construction with paste, and retrieval with read_html
  scraped.url <- read_html(paste0("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=",
                                  url_pwsid, "&year=", url_year))

  #set element address tags as objects
  scraped.city.tag <- "div+ table tr:nth-child(1) td:nth-child(2)"
  scraped.pwsid.tag <- "td tr:nth-child(1) td:nth-child(5)"
  scraped.ownership.tag <- "div+ table tr:nth-child(2) td:nth-child(4)"
  scraped.mgd.tag <- "th~ td+ td"

  #scrape data items
  scraped.city <- scraped.url %>% html_nodes(scraped.city.tag) %>% html_text()
  scraped.pwsid <- scraped.url %>% html_nodes(scraped.pwsid.tag) %>% html_text()
  scraped.ownership <- scraped.url %>% html_nodes(scraped.ownership.tag) %>%
    html_text()
  scraped.mgd <- scraped.url %>% html_nodes(scraped.mgd.tag) %>% html_text()
  #convert to dataframe
  scraped.withdrawals <- data.frame("Month" = c("Jan", "May", "Sep", "Feb", "Jun",
                                               "Oct", "Mar", "Jul", "Nov",
                                               "Apr", "Aug", "Dec"),
                                    "Year" = rep(url_year, 12),
                                    "Max_Withdrawals" = as.numeric(scraped.mgd))

  scraped.withdrawals <- scraped.withdrawals %>%
  mutate(City = !!scraped.city,
         PWSID = !!scraped.pwsid,
         Ownership = !!scraped.ownership,
         Date = my(paste(Month,"-",Year)))

  Sys.sleep(2)

  return(scraped.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

#used scrape.page() function to extract Durham values
durham.2015.withdrawals <- scrape.page("03-32-010", 2015)

#used ggplot() and geom_line() to create line plot
durham.2015.plot <- durham.2015.withdrawals %>%
  ggplot(aes(x=Date, y=Max_Withdrawals)) +
  geom_line() +
  labs(title=paste("Max Monthly Day Use for",Durham,"in 2015"),
       x = "Month",
       y = "Withdrawal (MGD)")
durham.2015.plot
```

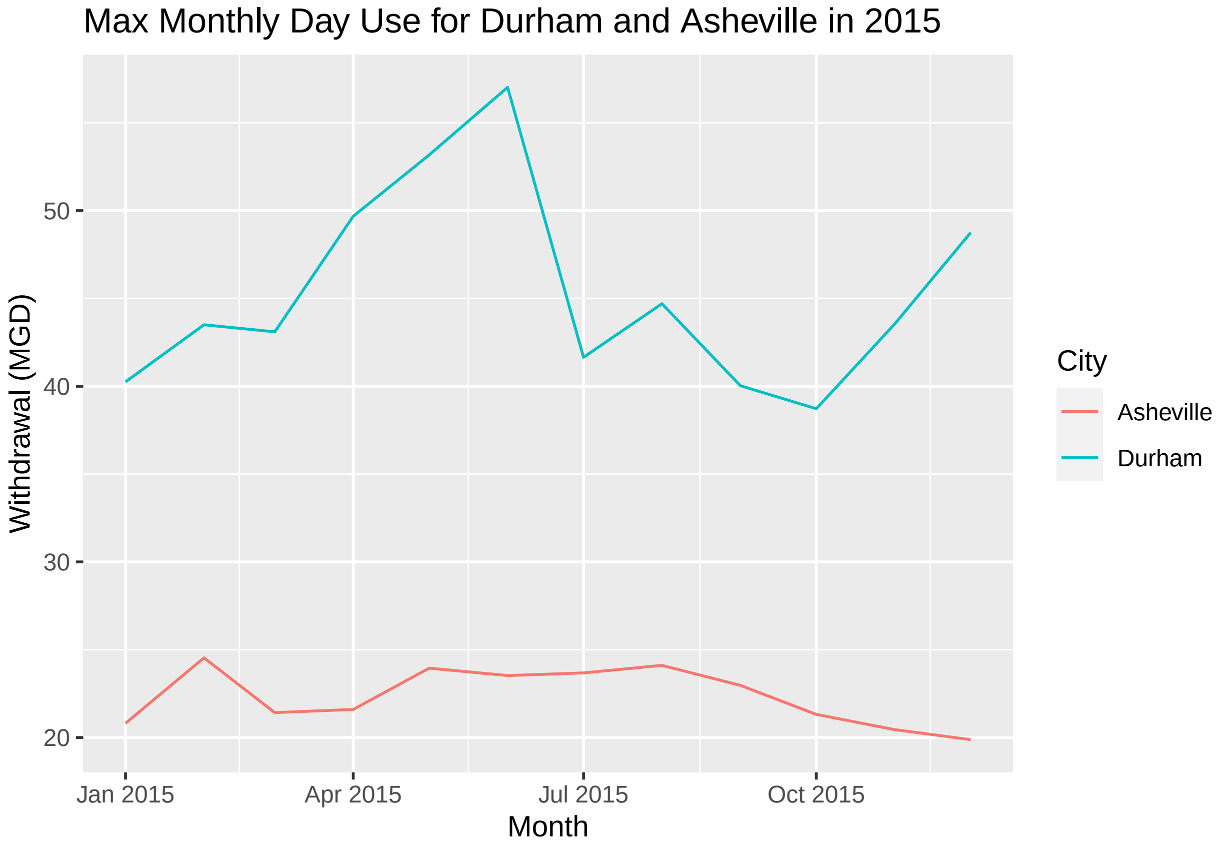


8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares Asheville's to Durham's water withdrawals.

```
#8

#used scrape.page() function to extract Asheville values
asheville.2015.withdrawals <- scrape.page("01-11-010", 2015)
asheville.2015.withdrawals
```

```
##    Month Year Max_Withdrawals      City     PWSID    Ownership       Date
## 1    Jan 2015           20.81 Asheville 01-11-010 Municipality 2015-01-01
## 2    May 2015           23.95 Asheville 01-11-010 Municipality 2015-05-01
## 3    Sep 2015           22.97 Asheville 01-11-010 Municipality 2015-09-01
## 4    Feb 2015           24.54 Asheville 01-11-010 Municipality 2015-02-01
## 5    Jun 2015           23.53 Asheville 01-11-010 Municipality 2015-06-01
## 6    Oct 2015           21.32 Asheville 01-11-010 Municipality 2015-10-01
## 7    Mar 2015           21.42 Asheville 01-11-010 Municipality 2015-03-01
## 8    Jul 2015           23.68 Asheville 01-11-010 Municipality 2015-07-01
## 9    Nov 2015           20.45 Asheville 01-11-010 Municipality 2015-11-01
## 10   Apr 2015           21.60 Asheville 01-11-010 Municipality 2015-04-01
## 11   Aug 2015           24.11 Asheville 01-11-010 Municipality 2015-08-01
## 12   Dec 2015           19.88 Asheville 01-11-010 Municipality 2015-12-01
```

```
#combine both Durham and Asheville data with bind_rows
withdrawals.2015 <- bind_rows(durham.2015.withdrawals, asheville.2015.withdrawals)

#used ggplot() and geom_line() to create line plot of combined datasets,
#differentiated between cities by setting color=City under geom_line aesthetics
withdrawals.2015.plot <- withdrawals.2015 %>%
  ggplot(aes(x=Date, y=Max_Withdrawals)) +
  geom_line(aes(color=City)) +
  labs(title=paste("Max Monthly Day Use for Durham and Asheville in 2015"),
       x = "Month",
       y = "Withdrawal (MGD)")
withdrawals.2015.plot
```



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.

```
#9

#used rep() to sequence the years from 2010 to 2021, and rep.int() to ensure the
#pwsid numbers are repeated the correct number of times
```

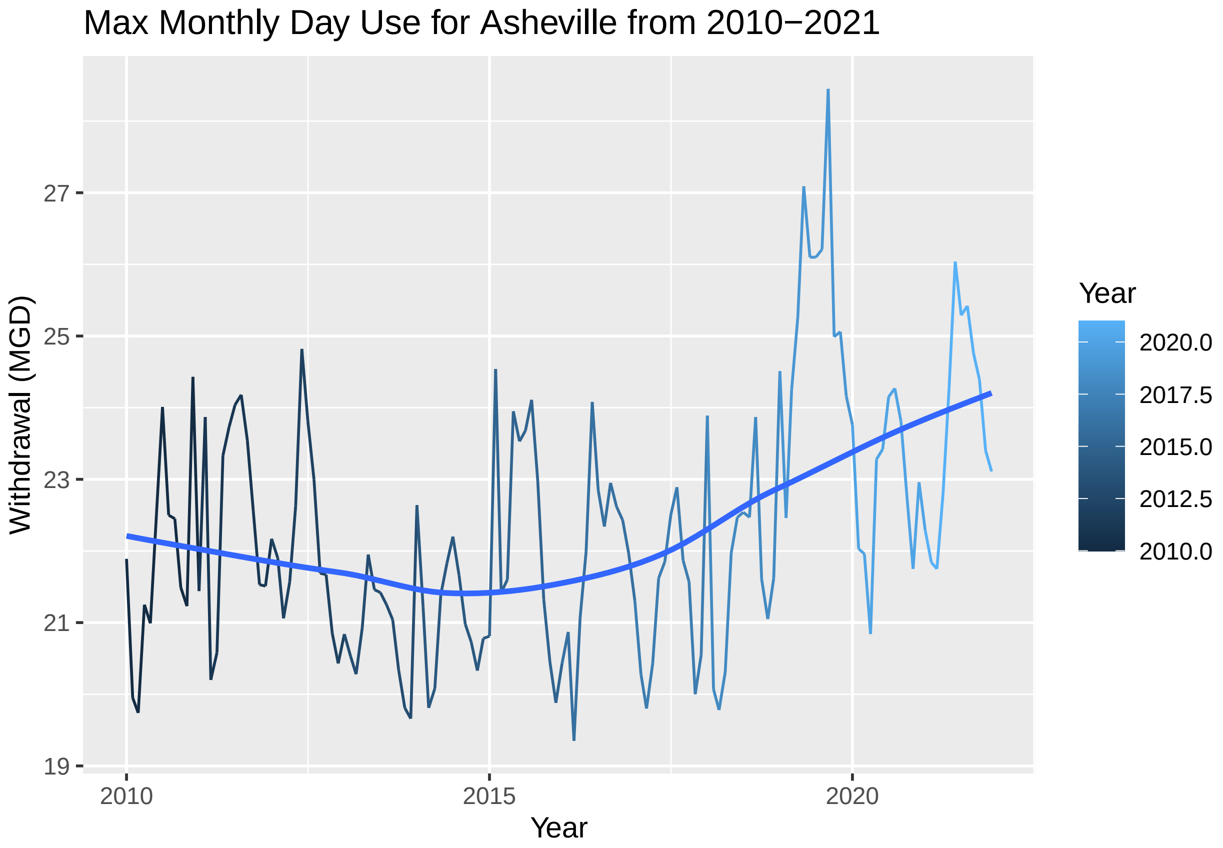
```r
asheville_years <- rep(2010:2021)
asheville_id <- rep.int("01-11-010", length(asheville_years))

#use map2 with the assigned objects in the correct function (scrape.page) order
asheville.2010.2021.dfs <- map2(asheville_id, asheville_years, scrape.page)

#bind all extracted datasets into one with bind_rows()
asheville.2010.2021.df <- bind_rows(asheville.2010.2021.dfs)
asheville.2010.2021.df
```

```
##     Month Year Max_Withdrawals      City     PWSID    Ownership       Date
## 1     Jan 2010           21.89 Asheville 01-11-010 Municipality 2010-01-01
## 2     May 2010           20.99 Asheville 01-11-010 Municipality 2010-05-01
## 3     Sep 2010           22.45 Asheville 01-11-010 Municipality 2010-09-01
## 4     Feb 2010           19.95 Asheville 01-11-010 Municipality 2010-02-01
## 5     Jun 2010           22.53 Asheville 01-11-010 Municipality 2010-06-01
## 6     Oct 2010           21.49 Asheville 01-11-010 Municipality 2010-10-01
## 7     Mar 2010           19.74 Asheville 01-11-010 Municipality 2010-03-01
## 8     Jul 2010           24.01 Asheville 01-11-010 Municipality 2010-07-01
## 9     Nov 2010           21.23 Asheville 01-11-010 Municipality 2010-11-01
## 10    Apr 2010           21.25 Asheville 01-11-010 Municipality 2010-04-01
## 11    Aug 2010           22.50 Asheville 01-11-010 Municipality 2010-08-01
## 12    Dec 2010           24.43 Asheville 01-11-010 Municipality 2010-12-01
## 13    Jan 2011           21.44 Asheville 01-11-010 Municipality 2011-01-01
## 14    May 2011           23.33 Asheville 01-11-010 Municipality 2011-05-01
## 15    Sep 2011           23.54 Asheville 01-11-010 Municipality 2011-09-01
## 16    Feb 2011           23.87 Asheville 01-11-010 Municipality 2011-02-01
## 17    Jun 2011           23.73 Asheville 01-11-010 Municipality 2011-06-01
## 18    Oct 2011           22.55 Asheville 01-11-010 Municipality 2011-10-01
## 19    Mar 2011           20.20 Asheville 01-11-010 Municipality 2011-03-01
## 20    Jul 2011           24.04 Asheville 01-11-010 Municipality 2011-07-01
## 21    Nov 2011           21.53 Asheville 01-11-010 Municipality 2011-11-01
## 22    Apr 2011           20.58 Asheville 01-11-010 Municipality 2011-04-01
## 23    Aug 2011           24.18 Asheville 01-11-010 Municipality 2011-08-01
## 24    Dec 2011           21.51 Asheville 01-11-010 Municipality 2011-12-01
## 25    Jan 2012           22.17 Asheville 01-11-010 Municipality 2012-01-01
## 26    May 2012           22.63 Asheville 01-11-010 Municipality 2012-05-01
## 27    Sep 2012           21.69 Asheville 01-11-010 Municipality 2012-09-01
## 28    Feb 2012           21.90 Asheville 01-11-010 Municipality 2012-02-01
## 29    Jun 2012           24.82 Asheville 01-11-010 Municipality 2012-06-01
## 30    Oct 2012           21.67 Asheville 01-11-010 Municipality 2012-10-01
## 31    Mar 2012           21.06 Asheville 01-11-010 Municipality 2012-03-01
## 32    Jul 2012           23.82 Asheville 01-11-010 Municipality 2012-07-01
## 33    Nov 2012           20.85 Asheville 01-11-010 Municipality 2012-11-01
## 34    Apr 2012           21.57 Asheville 01-11-010 Municipality 2012-04-01
## 35    Aug 2012           23.00 Asheville 01-11-010 Municipality 2012-08-01
## 36    Dec 2012           20.43 Asheville 01-11-010 Municipality 2012-12-01
## 37    Jan 2013           20.84 Asheville 01-11-010 Municipality 2013-01-01
## 38    May 2013           21.95 Asheville 01-11-010 Municipality 2013-05-01
## 39    Sep 2013           21.04 Asheville 01-11-010 Municipality 2013-09-01
## 40    Feb 2013           20.53 Asheville 01-11-010 Municipality 2013-02-01
## 41    Jun 2013           21.46 Asheville 01-11-010 Municipality 2013-06-01
## 42    Oct 2013           20.34 Asheville 01-11-010 Municipality 2013-10-01
## 43    Mar 2013           20.28 Asheville 01-11-010 Municipality 2013-03-01
## 44    Jul 2013           21.42 Asheville 01-11-010 Municipality 2013-07-01
## 45    Nov 2013           19.81 Asheville 01-11-010 Municipality 2013-11-01
## 46    Apr 2013           20.93 Asheville 01-11-010 Municipality 2013-04-01
## 47    Aug 2013           21.25 Asheville 01-11-010 Municipality 2013-08-01
## 48    Dec 2013           19.66 Asheville 01-11-010 Municipality 2013-12-01
## 49    Jan 2014           22.64 Asheville 01-11-010 Municipality 2014-01-01
## 50    May 2014           21.39 Asheville 01-11-010 Municipality 2014-05-01
## 51    Sep 2014           20.98 Asheville 01-11-010 Municipality 2014-09-01
## 52    Feb 2014           21.22 Asheville 01-11-010 Municipality 2014-02-01
## 53    Jun 2014           21.83 Asheville 01-11-010 Municipality 2014-06-01
## 54    Oct 2014           20.73 Asheville 01-11-010 Municipality 2014-10-01
## 55    Mar 2014           19.81 Asheville 01-11-010 Municipality 2014-03-01
## 56    Jul 2014           22.20 Asheville 01-11-010 Municipality 2014-07-01
## 57    Nov 2014           20.33 Asheville 01-11-010 Municipality 2014-11-01
## 58    Apr 2014           20.08 Asheville 01-11-010 Municipality 2014-04-01
## 59    Aug 2014           21.66 Asheville 01-11-010 Municipality 2014-08-01
## 60    Dec 2014           20.78 Asheville 01-11-010 Municipality 2014-12-01
## 61    Jan 2015           20.81 Asheville 01-11-010 Municipality 2015-01-01
## 62    May 2015           23.95 Asheville 01-11-010 Municipality 2015-05-01
## 63    Sep 2015           22.97 Asheville 01-11-010 Municipality 2015-09-01
## 64    Feb 2015           24.54 Asheville 01-11-010 Municipality 2015-02-01
## 65    Jun 2015           23.53 Asheville 01-11-010 Municipality 2015-06-01
## 66    Oct 2015           21.32 Asheville 01-11-010 Municipality 2015-10-01
## 67    Mar 2015           21.42 Asheville 01-11-010 Municipality 2015-03-01
## 68    Jul 2015           23.68 Asheville 01-11-010 Municipality 2015-07-01
## 69    Nov 2015           20.45 Asheville 01-11-010 Municipality 2015-11-01
## 70    Apr 2015           21.60 Asheville 01-11-010 Municipality 2015-04-01
## 71    Aug 2015           24.11 Asheville 01-11-010 Municipality 2015-08-01
## 72    Dec 2015           19.88 Asheville 01-11-010 Municipality 2015-12-01
## 73    Jan 2016           20.43 Asheville 01-11-010 Municipality 2016-01-01
## 74    May 2016           21.99 Asheville 01-11-010 Municipality 2016-05-01
## 75    Sep 2016           22.95 Asheville 01-11-010 Municipality 2016-09-01
## 76    Feb 2016           20.87 Asheville 01-11-010 Municipality 2016-02-01
## 77    Jun 2016           24.08 Asheville 01-11-010 Municipality 2016-06-01
## 78    Oct 2016           22.62 Asheville 01-11-010 Municipality 2016-10-01
## 79    Mar 2016           19.35 Asheville 01-11-010 Municipality 2016-03-01
## 80    Jul 2016           22.85 Asheville 01-11-010 Municipality 2016-07-01
## 81    Nov 2016           22.43 Asheville 01-11-010 Municipality 2016-11-01
## 82    Apr 2016           21.07 Asheville 01-11-010 Municipality 2016-04-01
## 83    Aug 2016           22.34 Asheville 01-11-010 Municipality 2016-08-01
## 84    Dec 2016           21.97 Asheville 01-11-010 Municipality 2016-12-01
## 85    Jan 2017           21.31 Asheville 01-11-010 Municipality 2017-01-01
## 86    May 2017           21.62 Asheville 01-11-010 Municipality 2017-05-01
## 87    Sep 2017           21.87 Asheville 01-11-010 Municipality 2017-09-01
## 88    Feb 2017           20.28 Asheville 01-11-010 Municipality 2017-02-01
## 89    Jun 2017           21.85 Asheville 01-11-010 Municipality 2017-06-01
## 90    Oct 2017           21.57 Asheville 01-11-010 Municipality 2017-10-01
## 91    Mar 2017           19.80 Asheville 01-11-010 Municipality 2017-03-01
## 92    Jul 2017           22.50 Asheville 01-11-010 Municipality 2017-07-01
## 93    Nov 2017           20.00 Asheville 01-11-010 Municipality 2017-11-01
## 94    Apr 2017           20.43 Asheville 01-11-010 Municipality 2017-04-01
## 95    Aug 2017           22.89 Asheville 01-11-010 Municipality 2017-08-01
## 96    Dec 2017           20.55 Asheville 01-11-010 Municipality 2017-12-01
## 97    Jan 2018           23.89 Asheville 01-11-010 Municipality 2018-01-01
## 98    May 2018           21.97 Asheville 01-11-010 Municipality 2018-05-01
## 99    Sep 2018           23.87 Asheville 01-11-010 Municipality 2018-09-01
## 100   Feb 2018           20.07 Asheville 01-11-010 Municipality 2018-02-01
## 101   Jun 2018           22.47 Asheville 01-11-010 Municipality 2018-06-01
## 102   Oct 2018           21.61 Asheville 01-11-010 Municipality 2018-10-01
## 103   Mar 2018           19.78 Asheville 01-11-010 Municipality 2018-03-01
## 104   Jul 2018           22.54 Asheville 01-11-010 Municipality 2018-07-01
## 105   Nov 2018           21.05 Asheville 01-11-010 Municipality 2018-11-01
## 106   Apr 2018           20.31 Asheville 01-11-010 Municipality 2018-04-01
## 107   Aug 2018           22.47 Asheville 01-11-010 Municipality 2018-08-01
## 108   Dec 2018           21.62 Asheville 01-11-010 Municipality 2018-12-01
## 109   Jan 2019           24.51 Asheville 01-11-010 Municipality 2019-01-01
## 110   May 2019           27.09 Asheville 01-11-010 Municipality 2019-05-01
## 111   Sep 2019           28.45 Asheville 01-11-010 Municipality 2019-09-01
## 112   Feb 2019           22.46 Asheville 01-11-010 Municipality 2019-02-01
## 113   Jun 2019           26.10 Asheville 01-11-010 Municipality 2019-06-01
## 114   Oct 2019           24.99 Asheville 01-11-010 Municipality 2019-10-01
## 115   Mar 2019           24.25 Asheville 01-11-010 Municipality 2019-03-01
## 116   Jul 2019           26.10 Asheville 01-11-010 Municipality 2019-07-01
## 117   Nov 2019           25.06 Asheville 01-11-010 Municipality 2019-11-01
## 118   Apr 2019           25.26 Asheville 01-11-010 Municipality 2019-04-01
## 119   Aug 2019           26.21 Asheville 01-11-010 Municipality 2019-08-01
## 120   Dec 2019           24.16 Asheville 01-11-010 Municipality 2019-12-01
## 121   Jan 2020           23.76 Asheville 01-11-010 Municipality 2020-01-01
## 122   May 2020           23.28 Asheville 01-11-010 Municipality 2020-05-01
## 123   Sep 2020           23.81 Asheville 01-11-010 Municipality 2020-09-01
## 124   Feb 2020           22.03 Asheville 01-11-010 Municipality 2020-02-01
## 125   Jun 2020           23.42 Asheville 01-11-010 Municipality 2020-06-01
## 126   Oct 2020           22.76 Asheville 01-11-010 Municipality 2020-10-01
## 127   Mar 2020           21.96 Asheville 01-11-010 Municipality 2020-03-01
## 128   Jul 2020           24.15 Asheville 01-11-010 Municipality 2020-07-01
## 129   Nov 2020           21.75 Asheville 01-11-010 Municipality 2020-11-01
## 130   Apr 2020           20.84 Asheville 01-11-010 Municipality 2020-04-01
## 131   Aug 2020           24.27 Asheville 01-11-010 Municipality 2020-08-01
## 132   Dec 2020           22.96 Asheville 01-11-010 Municipality 2020-12-01
## 133   Jan 2021           22.29 Asheville 01-11-010 Municipality 2021-01-01
## 134   May 2021           24.27 Asheville 01-11-010 Municipality 2021-05-01
## 135   Sep 2021           24.76 Asheville 01-11-010 Municipality 2021-09-01
## 136   Feb 2021           21.84 Asheville 01-11-010 Municipality 2021-02-01
## 137   Jun 2021           26.04 Asheville 01-11-010 Municipality 2021-06-01
## 138   Oct 2021           24.39 Asheville 01-11-010 Municipality 2021-10-01
## 139   Mar 2021           21.75 Asheville 01-11-010 Municipality 2021-03-01
## 140   Jul 2021           25.29 Asheville 01-11-010 Municipality 2021-07-01
## 141   Nov 2021           23.40 Asheville 01-11-010 Municipality 2021-11-01
## 142   Apr 2021           22.81 Asheville 01-11-010 Municipality 2021-04-01
## 143   Aug 2021           25.42 Asheville 01-11-010 Municipality 2021-08-01
## 144   Dec 2021           23.11 Asheville 01-11-010 Municipality 2021-12-01
```

```r
#plot all Asheville withdrawals and apply smoothed line with geom_smooth
asheville.withdrawals.plot <- asheville.2010.2021.df %>%
  ggplot(aes(x=Date, y=Max_Withdrawals, color=Year)) +
  geom_line() +
  geom_smooth(method="loess", se=FALSE) +
  labs(title=paste("Max Monthly Day Use for Asheville from 2010-2021"),
       x = "Year",
       y = "Withdrawal (MGD)")
asheville.withdrawals.plot
```

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

```
## Warning: The following aesthetics were dropped during statistical transformation: colour
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time? > Answer: Asheville has an upward trend in water usage after 2015. The upward trend may have been skewed by several large peaks around 2019, so running statistical analysis would still be helpful. >