SamanthaSedar_A10_DataScraping.Rmd

Samantha Sedar

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
library(tidyverse)
library(lubridate)
library(here); here()
```

```
## [1] "/home/guest/EDE_Fall2023"
```

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

# Set theme
mytheme <- theme_classic() +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
theme_set(mytheme)

here()
```

```
## [1] "/home/guest/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
#Fetch the web resources from the URL
lwsp <- read_html('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2022')
lwsp
```

```
## {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 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

#Scraping data using selector gadget

water_system_name <- lwsp %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()
water_system_name
```

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

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

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

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

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

```
MGD <- lwsp %>%
  html_nodes("th~ td+ td , th~ td+ td") %>%
  html_text()
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.

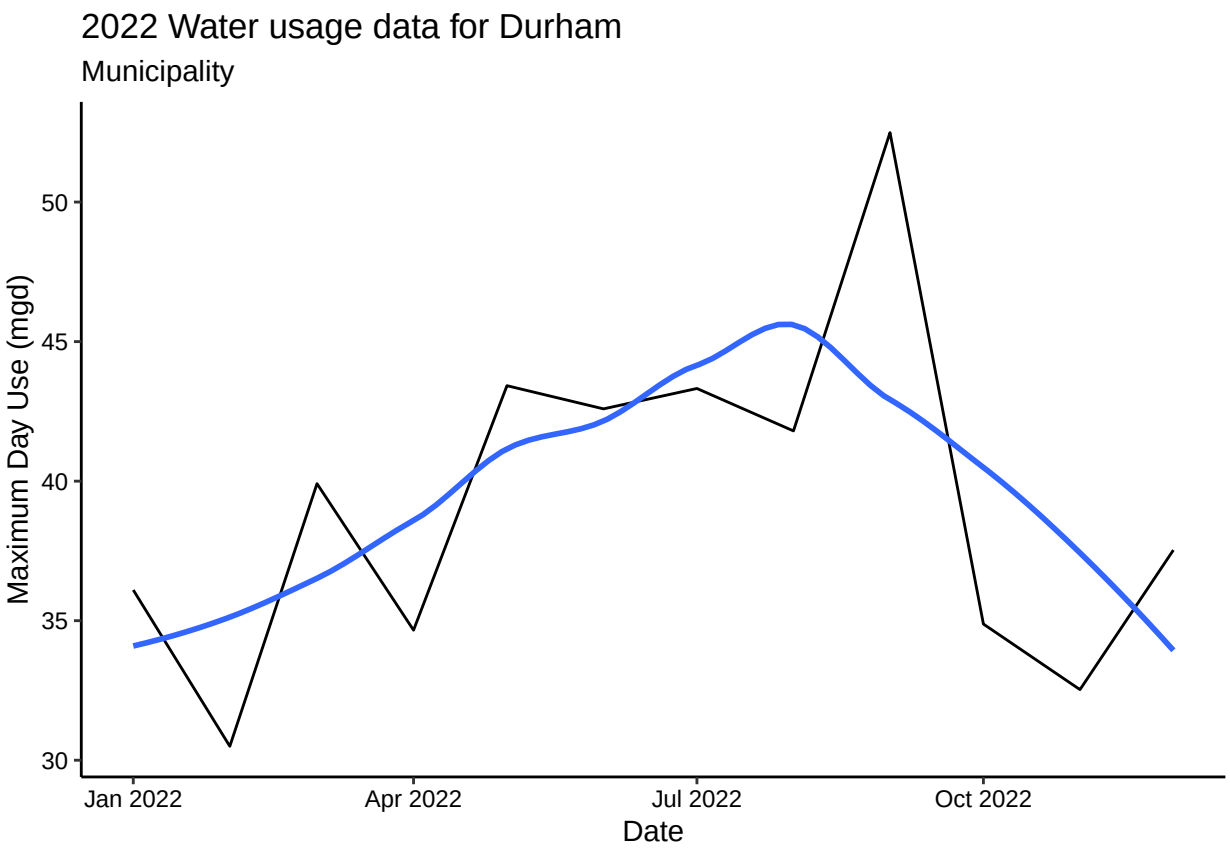
   NOTE: It's likely you won't be able to scrape the monthly widthrawal data in chronological order. You can overcome this by creating a month column manually assigning values in the order the data are scraped: "Jan", "May", "Sept", "Feb", etc... Or, you could scrape month values from the web page...

5. Create a line plot of the maximum daily withdrawals across the months for 2022

```
#4
#vector for specific order
months_order <- c(1, 5, 9, 2, 6, 10, 3, 7,11, 4, 8, 12)

df_lwsp_withdrawals <- data.frame("Month" = months_order,
                                  "Year" = rep(2022,12),
                                  "Maximum Day Use" = as.numeric(MGD))

#Modify the dataframe to include the other scraped variables as well as the date (as date object)
df_lwsp_withdrawals <- df_lwsp_withdrawals %>%
  mutate(Water_System_Name = !!water_system_name,
         PWSID = !!PWSID,
         Ownership = !!Ownership,
         Date = my(paste(Month,"-",Year)))
```

```r
#5
#Plot
ggplot(df_lwsp_withdrawals,aes(x=Date,y=Maximum.Day.Use)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2022 Water usage data for",water_system_name),
       subtitle = Ownership,
       y="Maximum Day Use (mgd)",
       x="Date")
```

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



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**.

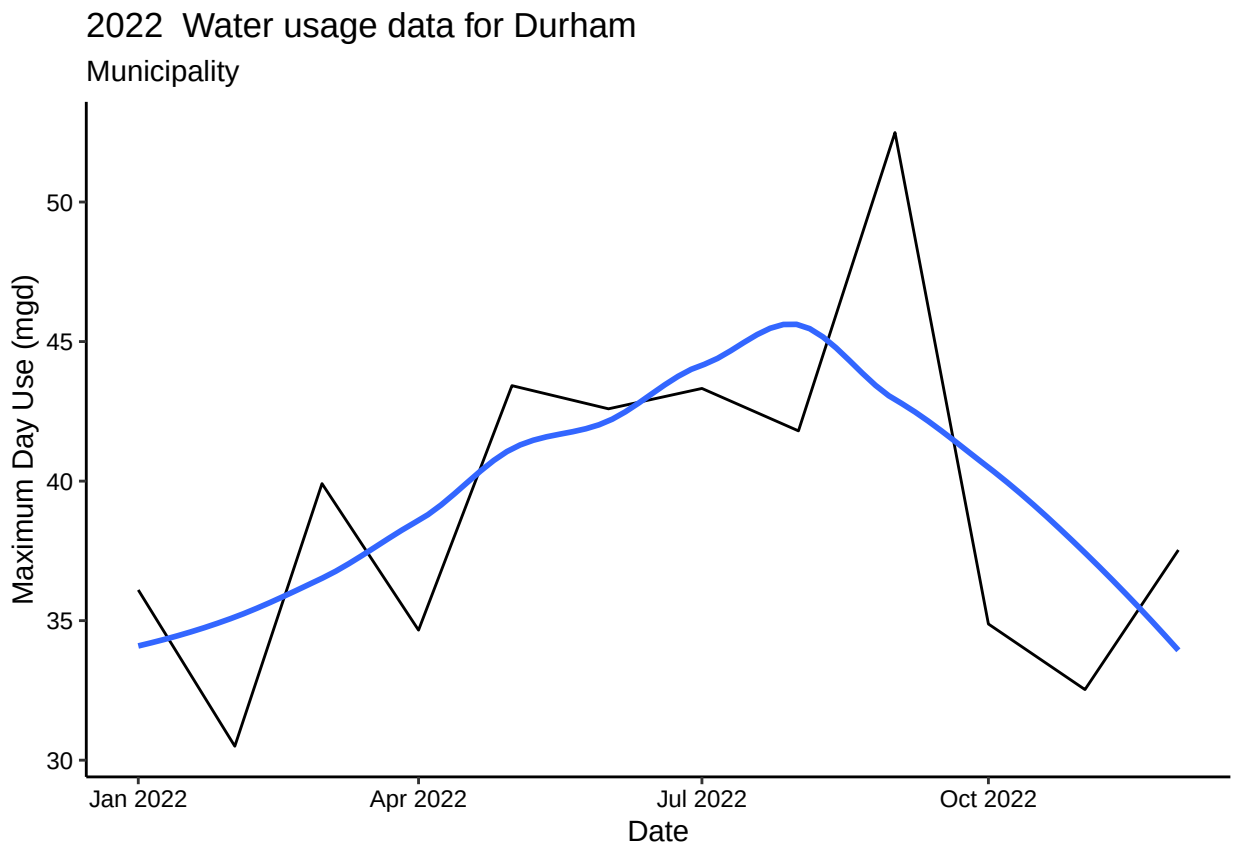
```r
#6.

#Construct the scraping web address, i.e. its URL
the.base.url <- 'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid='
the.pwsid <- '03-32-010'
the.year <- 2022
```

```
the.scrape.url <- paste0(the.base.url, the.pwsid, '&year=', the.year)
print(the.scrape.url)
```

```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2022"
```

```
#Retrieve the website contents
the.website <- read_html(the.scrape.url)

#Set the element address variables
the_water_system_name_tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'
the_PWSID_tag <- 'td tr:nth-child(1) td:nth-child(5)'
the_ownership_tag <- 'div+ table tr:nth-child(2) td:nth-child(4)'
the_mgd_tag <- 'th~ td+ td , th~ td+ td'

months_order <- c(1, 5, 9, 2, 6, 10, 3, 7,11, 4, 8, 12)

#Scrape the data items
water_system_name <- the.website %>% html_nodes(the_water_system_name_tag) %>% html_text()
PWSID <- the.website %>% html_nodes(the_PWSID_tag) %>% html_text()
Ownership <- the.website %>% html_nodes(the_ownership_tag) %>% html_text()
MGD <- the.website %>% html_nodes(the_mgd_tag) %>% html_text()

#Construct a dataframe from the scraped data

months_order <- c(1, 5, 9, 2, 6, 10, 3, 7,11, 4, 8, 12)
df_lwsp_withdrawals <- data.frame("Month" = months_order,
                                  "Year" = rep(the.year,12),
                                  "Maximum Day Use" = as.numeric(MGD)) %>%
  mutate(Water_System_Name = !!water_system_name,
         PWSID = !!PWSID,
         Ownership = !!Ownership,
         Date = my(paste(Month,"-",Year)))

#Plot
ggplot(df_lwsp_withdrawals,aes(x=Date,y=Maximum.Day.Use)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste(the.year," Water usage data for",water_system_name),
       subtitle = Ownership,
       y="Maximum Day Use (mgd)",
       x="Date")
```

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



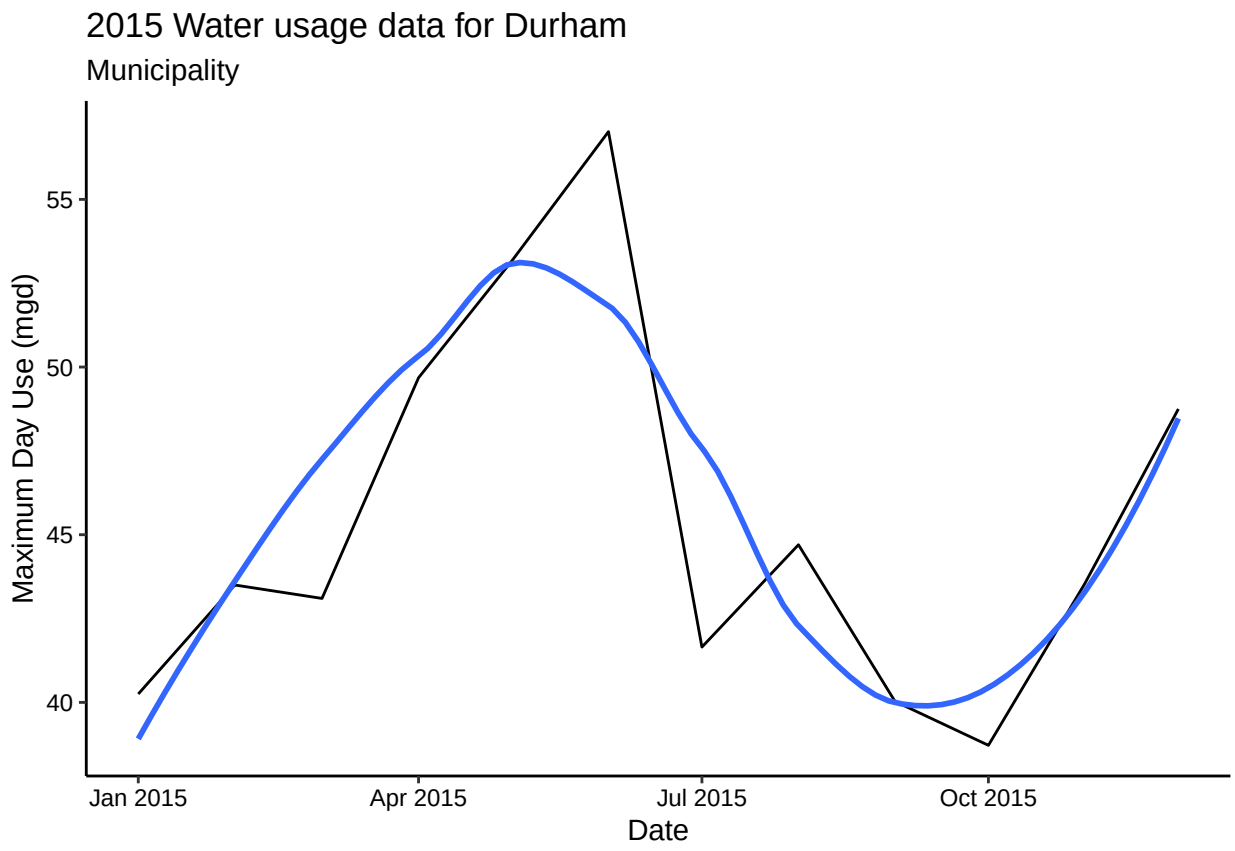

```
#Create our scraping function
scrape_it <- function(the.year, the.pwsid){

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

  #Set the element address variables
  the_water_system_name_tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'
  the_PWSID_tag <- 'td tr:nth-child(1) td:nth-child(5)'
  the_ownership_tag <- 'div+ table tr:nth-child(2) td:nth-child(4)'
  the_mgd_tag <- 'th~ td+ td , th~ td+ td'

  #Scrape the data items
  water_system_name <- the.website %>% html_nodes(the_water_system_name_tag) %>% html_text()
  PWSID <- the.website %>% html_nodes(the_PWSID_tag) %>% html_text()
  Ownership <- the.website %>% html_nodes(the_ownership_tag) %>% html_text()
  MGD <- the.website %>% html_nodes(the_mgd_tag) %>% html_text()

  #Convert to a dataframe
  months_order <- c(1, 5, 9, 2, 6, 10, 3, 7,11, 4, 8, 12)
  df_lwsp_withdrawals <- data.frame("Month" = months_order,
                            "Year" = rep(the.year,12),
                            "Maximum Day Use" = as.numeric(MGD)) %>%
    mutate(Water_System_Name = !!water_system_name,
           PWSID = !!PWSID,
```

```
           Ownership = !!Ownership,
          Date = my(paste(Month,"-",Year)))

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

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

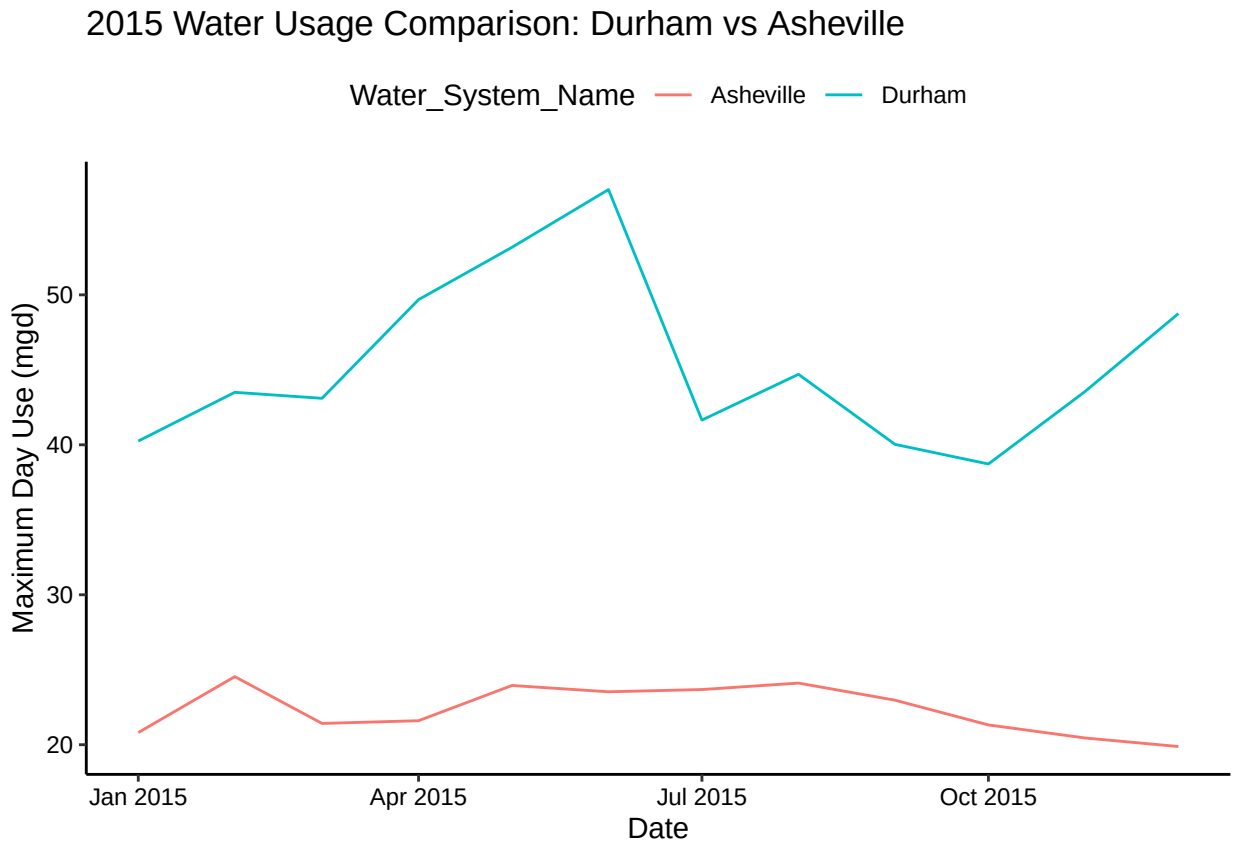
```
#7
#Using function

the_lwsp_df <- scrape_it(2015,'03-32-010')

ggplot(the_lwsp_df,aes(x=Date,y=Maximum.Day.Use)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2015 Water usage data for",water_system_name),
       subtitle = Ownership,
       y="Maximum Day Use (mgd)",
       x="Date")
```

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



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
#Using function
asheville_data <- scrape_it(2015, '01-11-010')
durham_data <- scrape_it(2015, '03-32-010')

#bind
combined_data <- rbind(asheville_data, durham_data)

# Plot
ggplot(combined_data, aes(x = Date, y = Maximum.Day.Use, color = Water_System_Name)) +
  geom_line() +
  labs(title = "2015 Water Usage Comparison: Durham vs Asheville",
       y = "Maximum Day Use (mgd)",
       x = "Date")
```

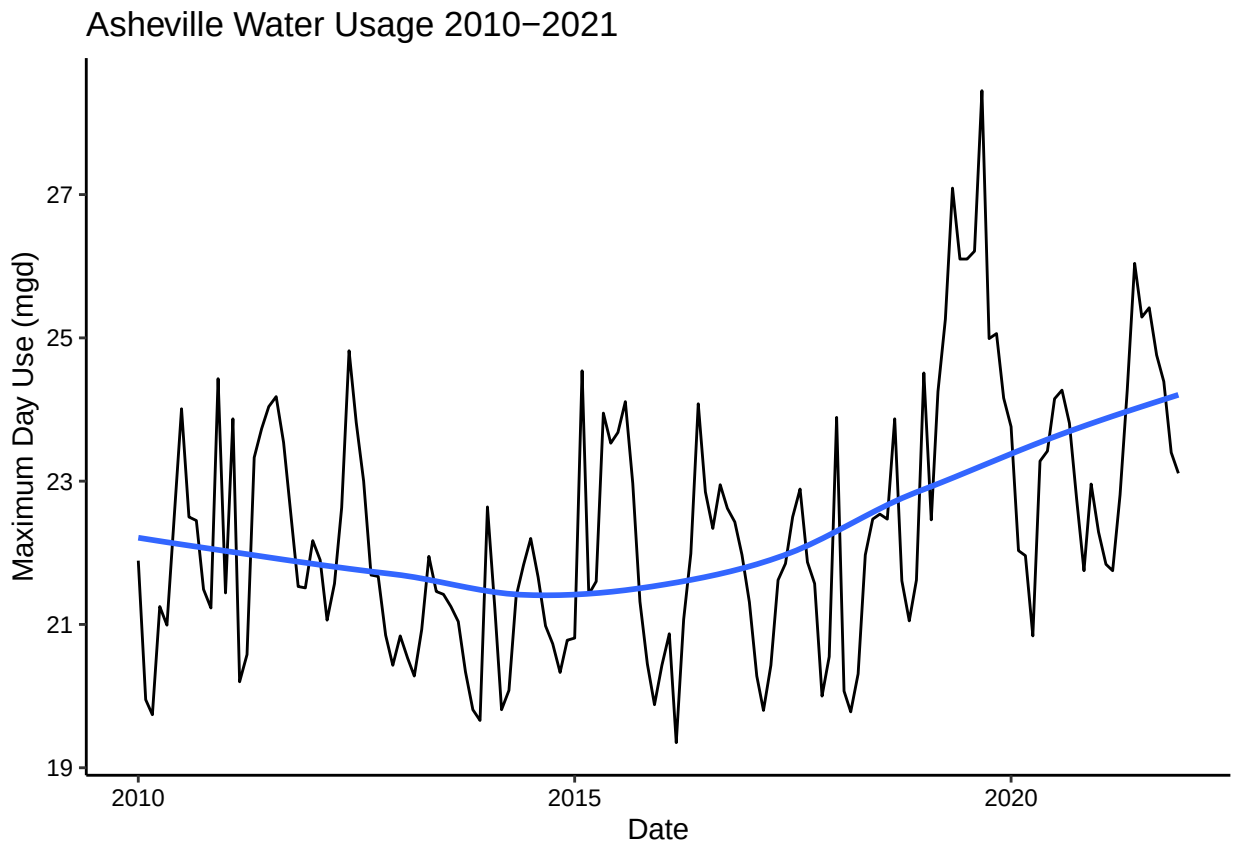


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2021.Add a smoothed line to the plot (method = 'loess').

   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.

```r
#9

the_years <- 2010:2021
asheville_facility_id <- '01-11-010'

facility_ids_asheville <- rep(asheville_facility_id, length(the_years))
dfs_asheville <- map2(the_years, facility_ids_asheville, scrape_it)
df_asheville_combined <- bind_rows(dfs_asheville, .id = "Year")

# Plot
ggplot(df_asheville_combined, aes(x = Date, y = Maximum.Day.Use)) +
  geom_line() +
  geom_smooth(method = 'loess', se = FALSE) +
  labs(title = "Asheville Water Usage 2010-2021",
       y = "Maximum Day Use (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? > Answer: > Yes, from the plot we can identify two trends. Asheville's water usage appears to vary by month within the year. Further, we see an overall gradual increase in water usage.