

Assignment 10: Data Scraping

Jordan Mullens

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on data scraping.

Directions

1. Rename this file `<JordanMullens>_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(rvest)
library(lubridate)
library(dplyr)
library(here); here()
```

```
## [1] "/home/guest/EDA-Spring2023"
```

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?pwdid=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
webpageWUDC <- read_html('https://www.ncwater.org/WUDC/app/LWSP/report.php?psid=03-32-010&year=2022')
webpageWUDC
```

```
## {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), with the first value being “36.1000”.

```
#3
water.system.name <- webpageWUDC %>%
  html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>%
  html_text()
water.system.name
```

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

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

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

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

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

```
MGD <- webpageWUDC %>%
  html_nodes('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 withdrawal 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 max daily withdrawals across the months for 2022

```
#4
df_MGD <- data.frame(Month = c(1,5,9,2,6,10,3,7,11,4,8,12),
                     Year = rep(2022,12),
                     Maximum_Day_Use = as.numeric(MGD))

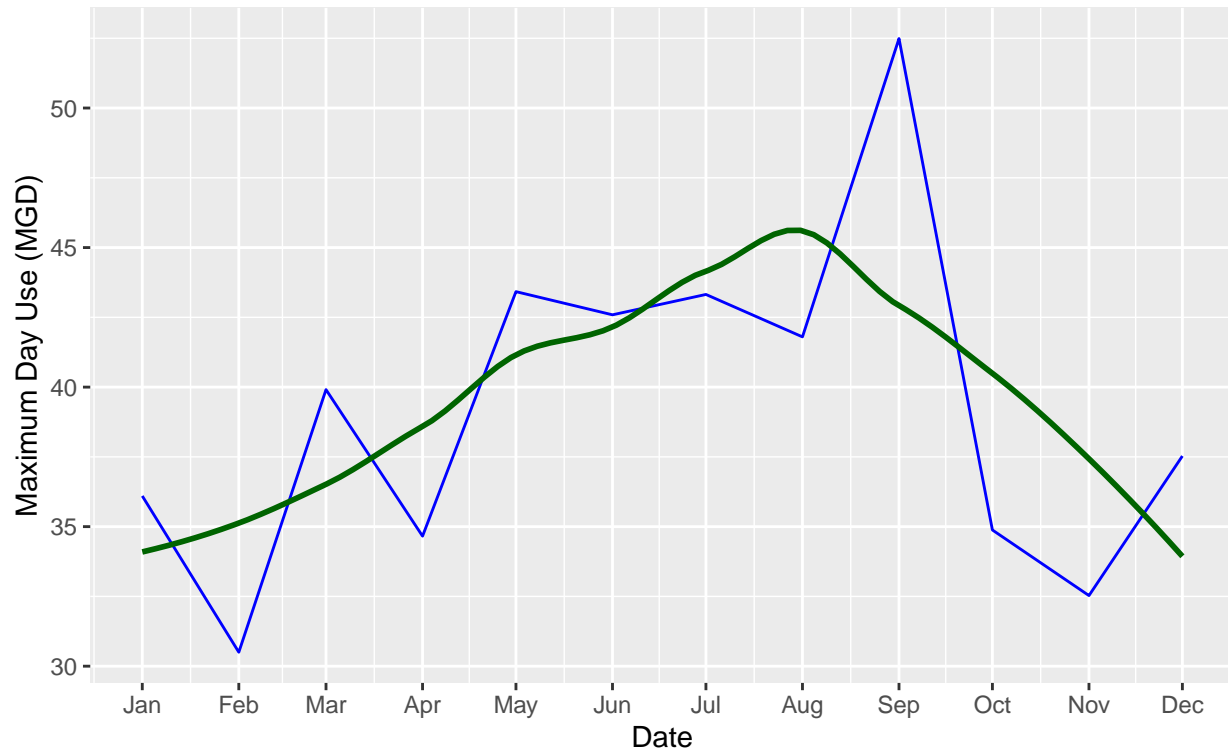
#Modify the dataframe to include the facility name and type as well as the date (as date object)
df_MGD <- df_MGD %>%
  mutate(Water_System = !!water.system.name,
         PWSID = !!PWSID,
         Ownership = !!ownership,
         Date = my(paste(Month,"-",Year)))

#5
ggplot(df_MGD,aes(x=Date,y=Maximum_Day_Use)) +
  geom_line(color="blue") +
  geom_smooth(method="loess",se=FALSE, color="darkgreen") +
  scale_x_date(date_breaks = "1 month", date_labels = "%b") +
  labs(title = paste("2022 Maximum Daily Water Usage Data for",PWSID),
       subtitle = water.system.name,
       y="Maximum Day Use (MGD)",
       x="Date")
```

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

2022 Maximum Daily Water Usage Data for 03-32-010

Durham



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.
#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
```

```
#Create our scraping function
scrape.MGD <- function(the_pwsid, the_year){

  #Retrieve the website contents
  the_website <- read_html(paste0(the_base_url,
                                   the_pwsid, '&year=', the_year))

  #Set the element address variables (determined in the previous step)
  water.system.tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'
  PWSID.tag <- 'td tr:nth-child(1) td:nth-child(5)'
  ownership.tag <- 'div+ table tr:nth-child(2) td:nth-child(4)'
  MGD.tag <- 'th~ td+ td'

  #Scrape the data items
  water.system.name <- the_website %>% html_nodes(water.system.tag) %>% html_text()
```

```

PWSID <- the_website %>% html_nodes(PWSID.tag) %>% html_text()
ownership <- the_website %>% html_nodes(ownership.tag) %>% html_text()
MGD <- the_website %>% html_nodes(MGD.tag) %>% html_text()

#Convert to a dataframe
df_MGD.f <- data.frame("Month" = c(1,5,9,2,6,10,3,7,11,4,8,12),
                      "Year" = rep(the_year,12),
                      "Maximum_Day_Use" = as.numeric(MGD)) %>%
  mutate(Water_System = !!water.system.name,
         PWSID = !!PWSID,
         Ownership_Type = !!ownership,
         Date = my(paste(Month,"-",Year)))

#Return the dataframe
return(df_MGD.f)
}

```

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

```

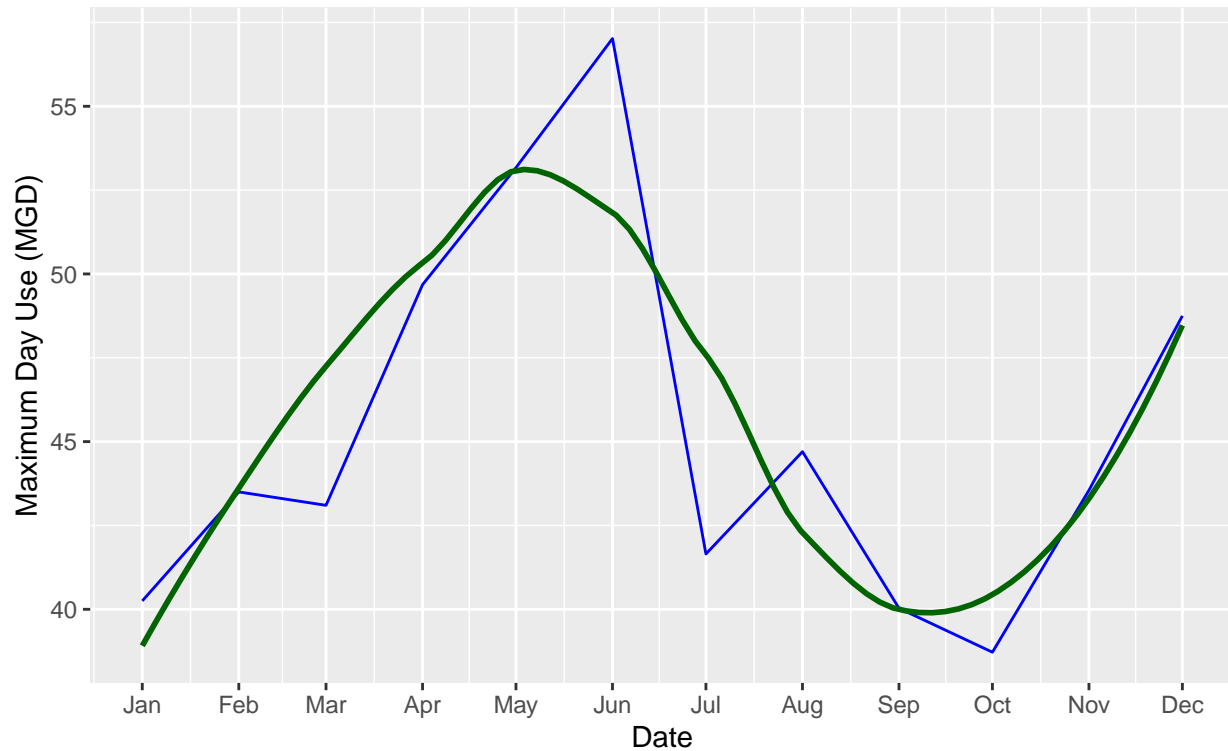
#7
Durham.2015.MGD <- scrape.MGD('03-32-010', 2015)
view(Durham.2015.MGD)

#plotting Durham 2015 Data
Durham.2015.plot <- ggplot(Durham.2015.MGD,aes(x=Date,y=Maximum_Day_Use)) +
  geom_line(color="blue") +
  geom_smooth(method="loess",se=FALSE, color="darkgreen") +
  scale_x_date(date_breaks = "1 month", date_labels = "%b") +
  labs(title = paste("2015 Maximum Daily Water Usage Data for",PWSID),
       subtitle = water.system.name,
       y="Maximum Day Use (MGD)",
       x="Date")
print(Durham.2015.plot)

```

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

2015 Maximum Daily Water Usage Data for 03-32-010 Durham



- 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
Asheville.2015.MGD <- scrape.MGD('01-11-010', 2015)
view(Asheville.2015.MGD)

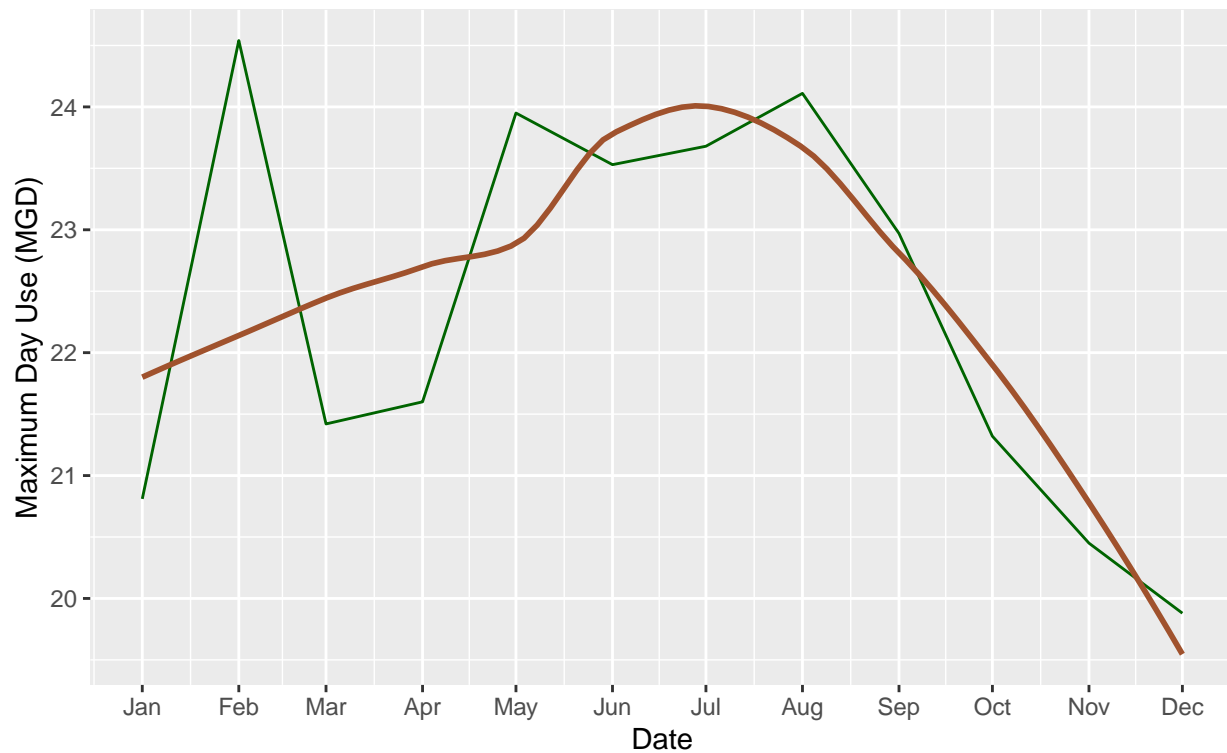
#plotting Asheville info to check the data
Asheville.2015.plot <- ggplot(Asheville.2015.MGD,aes(x=Date,y=Maximum_Day_Use)) +
  geom_line(color="darkgreen") +
  geom_smooth(method="loess",se=FALSE, color="sienna") +
  scale_x_date(date_breaks = "1 month", date_labels = "%b") +
  labs(title = paste("2015 Maximum Daily Water Usage Data for",PWSID),
        subtitle = water.system.name,
        y="Maximum Day Use (MGD)",
        x="Date")

print(Asheville.2015.plot)
```

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

2015 Maximum Daily Water Usage Data for 03-32-010

Durham



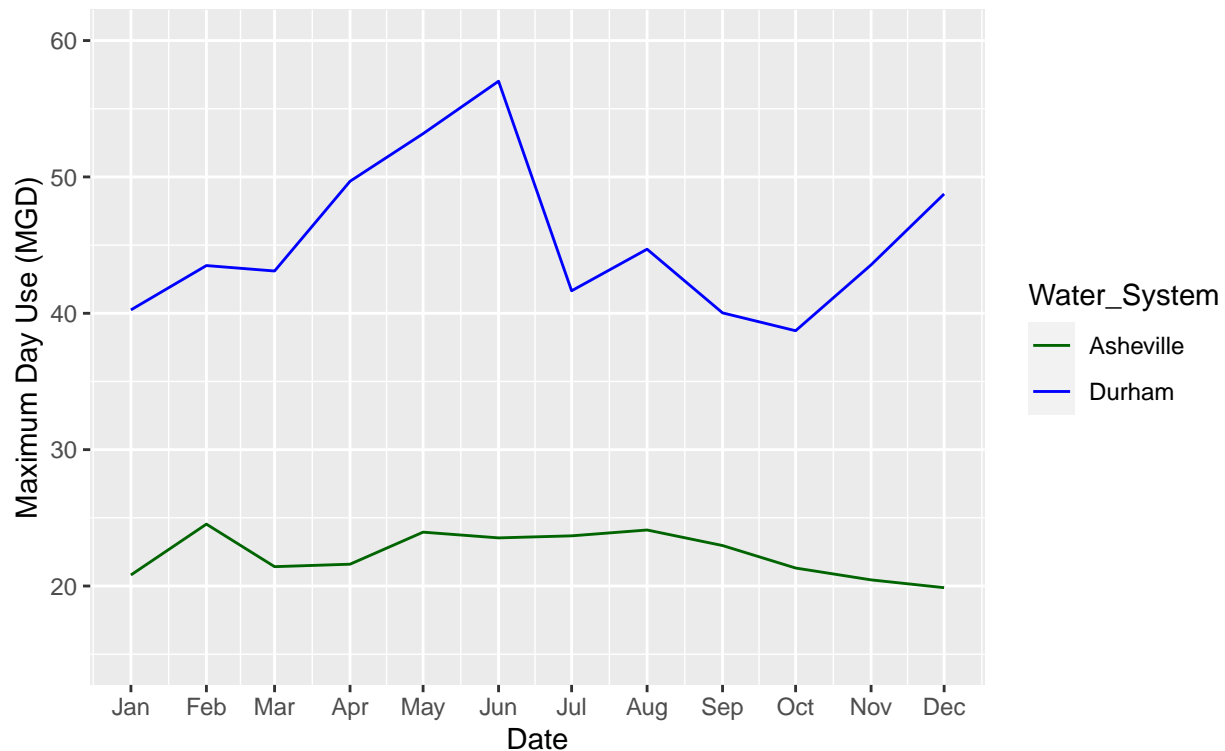
```
#combining Durham 2015 and Asheville 2015 data
Ash.Dur.2015.data <- rbind(Durham.2015.MGD, Asheville.2015.MGD)

Ash.Dur.2015.plot <- ggplot(Ash.Dur.2015.data,
                             aes(x=Date,
                                 y=Maximum_Day_Use,
                                 color= Water_System)) +

  geom_line() +
  ylim(15, 60) +
  scale_color_manual(values=c('darkgreen','blue')) +
  scale_x_date(date_breaks = "1 month", date_labels = "%b") +
  labs(title = paste("2015 Maximum Daily Water Usage Data for",PWSID),
       subtitle = water.system.name,
       y="Maximum Day Use (MGD)",
       x="Date")
print(Ash.Dur.2015.plot)
```

2015 Maximum Daily Water Usage Data for 03–32–010

Durham



- 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 "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
#Set the inputs to scrape years 2010 to 2021 for Asheville
the_years = rep(2010:2021)

#Use lapply to apply the scrape function
the_Ashville_dfs <- lapply(X = the_years,
                          FUN = scrape.MGD,
                          the_pwsid='01-11-010')

Asheville.11.yr <- bind_rows(the_Ashville_dfs)

Asheville.11.yr.plot <- ggplot(Asheville.11.yr, aes(x=Date, y=Maximum_Day_Use)) +
  geom_line(color="blue") +
  geom_smooth(method="loess", se=FALSE, color="darkgreen") +
  scale_x_date(date_breaks = "1 year", date_labels = "%Y") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = paste("2010 to 2021 Maximum Daily Water Usage Data for Asheville"),
```

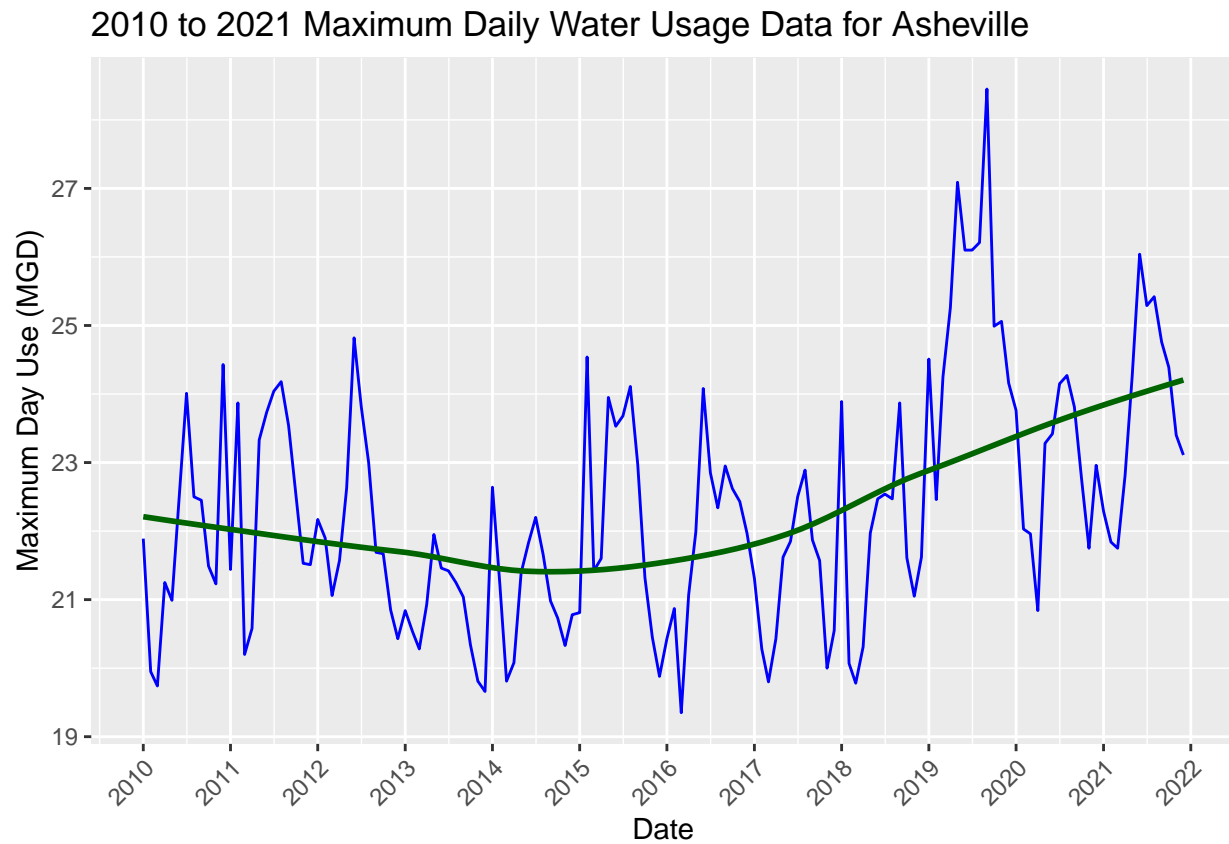


```

y="Maximum Day Use (MGD)",
x="Date")
print(Asheville.11yr.plot)

```

```
## '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? According to the graph, the water usage decreased from 2010 to 2015 and increased from 2015 to 2021. It is unclear whether these changes are significant or whether the water usage will continue to increase beyond 2021.