

Assignment 10: Data Scraping

Wynona Curaming

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(viridis);library(here);library(rvest)
library(dataRetrieval);library(tidycensus);library(ggplot2);library(dplyr)

getwd()
```

```
## [1] "/home/guest/ENV 872/EDA_Spring2024"
```

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

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

## [1] "Durham"

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

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

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

## [1] "Municipality"

max_day_use_monthly<-Water_URL%>%
  html_nodes("th~ td+ td")%>%
  html_text()
max_day_use_monthly

## [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"

max_day_use_monthly<-as.numeric(max_day_use_monthly)
class(max_day_use_monthly)

## [1] "numeric"
```

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 maximum daily withdrawals across the months for 2022

```
#4
#making a df
water_df<-data.frame(water_system_name,
                      PWSID,
                      "Ownership"=ownership,
                      "Max_day_use_monthly"=max_day_use_monthly)

#adding month & date columns
month<- c(1,5,9,2,6,10,3,7,11,4,8,12)
year<-c(2022)

water_df['Month']=month
water_df
```

##	water_system_name	PWSID	Ownership	Max_day_use_monthly	Month
## 1	Durham	03-32-010	Municipality	36.10	1
## 2	Durham	03-32-010	Municipality	43.42	5
## 3	Durham	03-32-010	Municipality	52.49	9
## 4	Durham	03-32-010	Municipality	30.50	2
## 5	Durham	03-32-010	Municipality	42.59	6
## 6	Durham	03-32-010	Municipality	34.88	10
## 7	Durham	03-32-010	Municipality	39.91	3
## 8	Durham	03-32-010	Municipality	43.32	7
## 9	Durham	03-32-010	Municipality	32.53	11
## 10	Durham	03-32-010	Municipality	34.66	4
## 11	Durham	03-32-010	Municipality	41.80	8
## 12	Durham	03-32-010	Municipality	37.53	12

```
water_df['Date']= my(paste(month, year))

water_df
```

##	water_system_name	PWSID	Ownership	Max_day_use_monthly	Month
## 1	Durham	03-32-010	Municipality	36.10	1
## 2	Durham	03-32-010	Municipality	43.42	5
## 3	Durham	03-32-010	Municipality	52.49	9
## 4	Durham	03-32-010	Municipality	30.50	2
## 5	Durham	03-32-010	Municipality	42.59	6
## 6	Durham	03-32-010	Municipality	34.88	10
## 7	Durham	03-32-010	Municipality	39.91	3
## 8	Durham	03-32-010	Municipality	43.32	7
## 9	Durham	03-32-010	Municipality	32.53	11
## 10	Durham	03-32-010	Municipality	34.66	4
## 11	Durham	03-32-010	Municipality	41.80	8
## 12	Durham	03-32-010	Municipality	37.53	12

##	Date
## 1	2022-01-01
## 2	2022-05-01
## 3	2022-09-01
## 4	2022-02-01
## 5	2022-06-01

```
## 6 2022-10-01
## 7 2022-03-01
## 8 2022-07-01
## 9 2022-11-01
## 10 2022-04-01
## 11 2022-08-01
## 12 2022-12-01
```

```
#reordering for chronological order
```

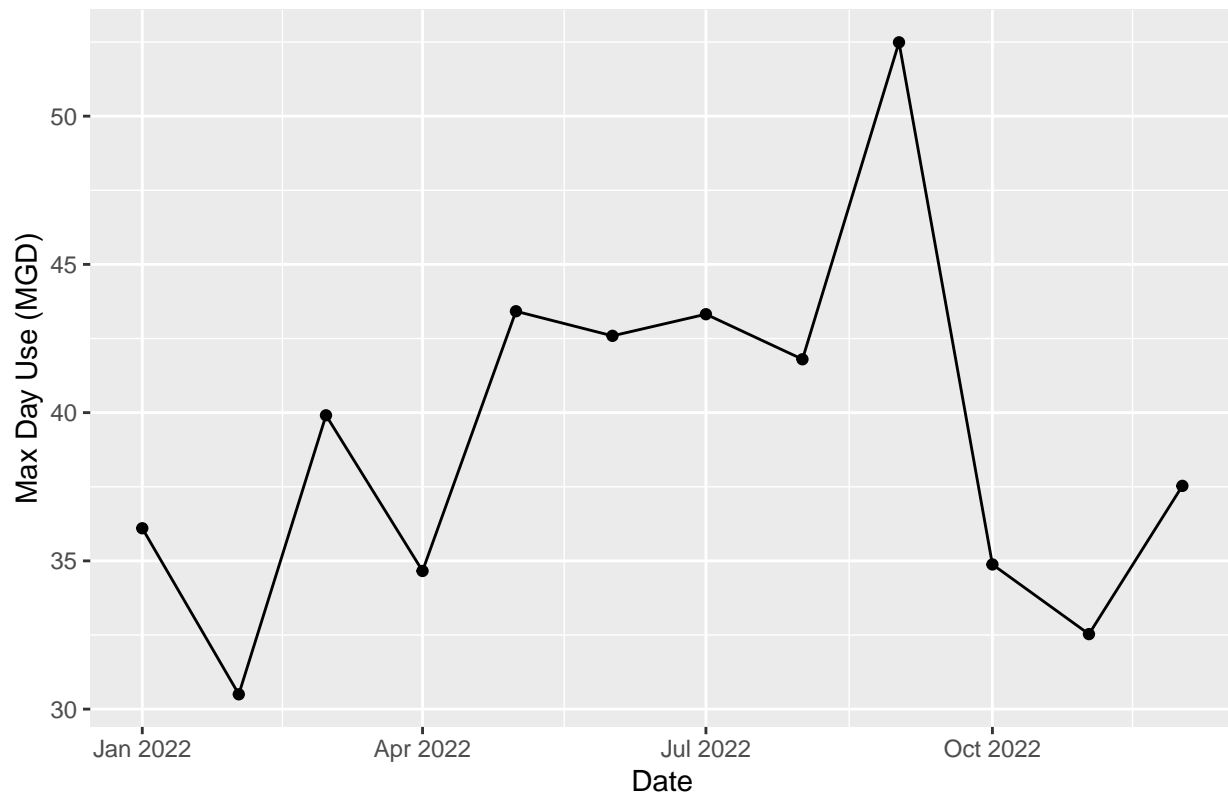
```
water_chronological_df<- water_df[order(water_df$Month),]
water_chronological_df
```

```
##      water_system_name      PWSID      Ownership Max_day_use_monthly Month
## 1      Durham 03-32-010 Municipality      36.10      1
## 4      Durham 03-32-010 Municipality      30.50      2
## 7      Durham 03-32-010 Municipality      39.91      3
## 10     Durham 03-32-010 Municipality      34.66      4
## 2      Durham 03-32-010 Municipality      43.42      5
## 5      Durham 03-32-010 Municipality      42.59      6
## 8      Durham 03-32-010 Municipality      43.32      7
## 11     Durham 03-32-010 Municipality      41.80      8
## 3      Durham 03-32-010 Municipality      52.49      9
## 6      Durham 03-32-010 Municipality      34.88     10
## 9      Durham 03-32-010 Municipality      32.53     11
## 12     Durham 03-32-010 Municipality      37.53     12
##      Date
## 1 2022-01-01
## 4 2022-02-01
## 7 2022-03-01
## 10 2022-04-01
## 2 2022-05-01
## 5 2022-06-01
## 8 2022-07-01
## 11 2022-08-01
## 3 2022-09-01
## 6 2022-10-01
## 9 2022-11-01
## 12 2022-12-01
```

```
#5
```

```
ggplot(aes(x=Date, y= Max_day_use_monthly), data=water_chronological_df)+geom_point()+geom_line()+
  labs(title="Max Day Use for Durham in 2022", y="Max Day Use (MGD)")
```

Max Day Use for Durham in 2022



- 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.

```
scrape.it<-function(the_year, the_facility){
  base_URL<-'https://www.ncwater.org/WUDC/app/LWSP/report.php?'
  the_scrape_url <- paste0(base_URL,'pwsid=',the_facility,'&year=', the_year)
  print(the_scrape_url)
  the_facility<-'03-32-010'
  the_website <- read_html(the_scrape_url)

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

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

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

  ownership_anyyear<-the_website%>%
    html_nodes("div+ table tr:nth-child(2) td:nth-child(4)")%>%
```

```

    html_text()
ownership

max_day_use_monthly_anyyear<-the_website%>%
  html_nodes("th~ td+ td")%>%
  html_text()

df_water<- data.frame("Month" = the_month,
                      "Year" = the_year,
                      "Max_Day_Use" = as.numeric(max_day_use_monthly_anyyear)) %>%
  mutate(Water_System_Name = !!water_system_name_anyyear,
         PWSID = !!PWSID_anyyear,
         Ownership = !!ownership_anyyear,
         Date = my(paste(Month,"-",Year)))
return(df_water)
}

```

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

```

#7
max_daily_withdrawals_Durham_2015_df <- scrape.it(2015,'03-32-010')

## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?psid=03-32-010&year=2015"
max_daily_withdrawals_Durham_2015_df

```

##	Month	Year	Max_Day_Use	Water_System_Name	PWSID	Ownership	Date
## 1	1	2015	40.25	Durham	03-32-010	Municipality	2015-01-01
## 2	5	2015	53.17	Durham	03-32-010	Municipality	2015-05-01
## 3	9	2015	40.03	Durham	03-32-010	Municipality	2015-09-01
## 4	2	2015	43.50	Durham	03-32-010	Municipality	2015-02-01
## 5	6	2015	57.02	Durham	03-32-010	Municipality	2015-06-01
## 6	10	2015	38.72	Durham	03-32-010	Municipality	2015-10-01
## 7	3	2015	43.10	Durham	03-32-010	Municipality	2015-03-01
## 8	7	2015	41.65	Durham	03-32-010	Municipality	2015-07-01
## 9	11	2015	43.55	Durham	03-32-010	Municipality	2015-11-01
## 10	4	2015	49.68	Durham	03-32-010	Municipality	2015-04-01
## 11	8	2015	44.70	Durham	03-32-010	Municipality	2015-08-01
## 12	12	2015	48.75	Durham	03-32-010	Municipality	2015-12-01

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
max_daily_withdrawals_Ashville_2015_df <- scrape.it(2015,'01-11-010')

## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?psid=01-11-010&year=2015"
max_daily_withdrawals_Ashville_2015_df

```

##	Month	Year	Max_Day_Use	Water_System_Name	PWSID	Ownership	Date
## 1	1	2015	20.81	Asheville	01-11-010	Municipality	2015-01-01
## 2	5	2015	23.95	Asheville	01-11-010	Municipality	2015-05-01
## 3	9	2015	22.97	Asheville	01-11-010	Municipality	2015-09-01
## 4	2	2015	24.54	Asheville	01-11-010	Municipality	2015-02-01

```
## 5      6 2015      23.53      Asheville 01-11-010 Municipality 2015-06-01
## 6     10 2015      21.32      Asheville 01-11-010 Municipality 2015-10-01
## 7      3 2015      21.42      Asheville 01-11-010 Municipality 2015-03-01
## 8      7 2015      23.68      Asheville 01-11-010 Municipality 2015-07-01
## 9     11 2015      20.45      Asheville 01-11-010 Municipality 2015-11-01
## 10     4 2015      21.60      Asheville 01-11-010 Municipality 2015-04-01
## 11     8 2015      24.11      Asheville 01-11-010 Municipality 2015-08-01
## 12    12 2015      19.88      Asheville 01-11-010 Municipality 2015-12-01
```

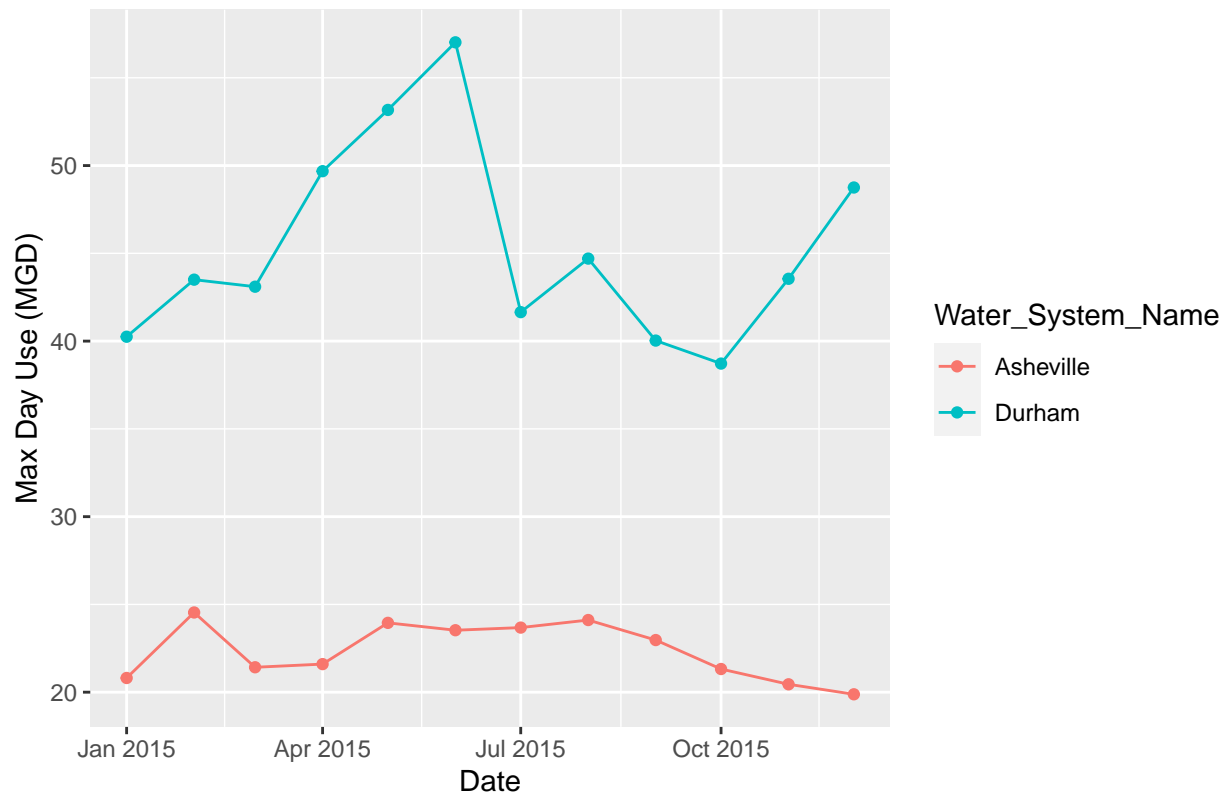
```
max_daily_withdrawals_Asheville_Durham_2015_df<-
  bind_rows(max_daily_withdrawals_Asheville_2015_df,
            max_daily_withdrawals_Durham_2015_df)
```

```
max_daily_withdrawals_Asheville_Durham_2015_df
```

##	Month	Year	Max_Day_Use	Water_System_Name	PWSID	Ownership	Date
## 1	1	2015	20.81	Asheville	01-11-010	Municipality	2015-01-01
## 2	5	2015	23.95	Asheville	01-11-010	Municipality	2015-05-01
## 3	9	2015	22.97	Asheville	01-11-010	Municipality	2015-09-01
## 4	2	2015	24.54	Asheville	01-11-010	Municipality	2015-02-01
## 5	6	2015	23.53	Asheville	01-11-010	Municipality	2015-06-01
## 6	10	2015	21.32	Asheville	01-11-010	Municipality	2015-10-01
## 7	3	2015	21.42	Asheville	01-11-010	Municipality	2015-03-01
## 8	7	2015	23.68	Asheville	01-11-010	Municipality	2015-07-01
## 9	11	2015	20.45	Asheville	01-11-010	Municipality	2015-11-01
## 10	4	2015	21.60	Asheville	01-11-010	Municipality	2015-04-01
## 11	8	2015	24.11	Asheville	01-11-010	Municipality	2015-08-01
## 12	12	2015	19.88	Asheville	01-11-010	Municipality	2015-12-01
## 13	1	2015	40.25	Durham	03-32-010	Municipality	2015-01-01
## 14	5	2015	53.17	Durham	03-32-010	Municipality	2015-05-01
## 15	9	2015	40.03	Durham	03-32-010	Municipality	2015-09-01
## 16	2	2015	43.50	Durham	03-32-010	Municipality	2015-02-01
## 17	6	2015	57.02	Durham	03-32-010	Municipality	2015-06-01
## 18	10	2015	38.72	Durham	03-32-010	Municipality	2015-10-01
## 19	3	2015	43.10	Durham	03-32-010	Municipality	2015-03-01
## 20	7	2015	41.65	Durham	03-32-010	Municipality	2015-07-01
## 21	11	2015	43.55	Durham	03-32-010	Municipality	2015-11-01
## 22	4	2015	49.68	Durham	03-32-010	Municipality	2015-04-01
## 23	8	2015	44.70	Durham	03-32-010	Municipality	2015-08-01
## 24	12	2015	48.75	Durham	03-32-010	Municipality	2015-12-01

```
ggplot(max_daily_withdrawals_Asheville_Durham_2015_df, aes(x=Date,
                                                            y=Max_Day_Use, color=Water_System_Name))+
  geom_point()+geom_line()+
  labs(title="Max Day Use for Asheville vs. Durham", y="Max Day Use (MGD)")
```

Max Day Use for Asheville vs. 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 "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
#Mapping out the data from 2010-2021 for Asheville (Quicker way)
water_df_2010_to_2021_v2<-map2(seq(2010,2021), '01-11-010', scrape.it)

## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2010"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2011"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2012"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2013"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2014"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2015"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2016"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2017"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2018"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2019"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2020"
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2021"

water_df_2010_to_2021_v2<-bind_rows(water_df_2010_to_2021_v2)

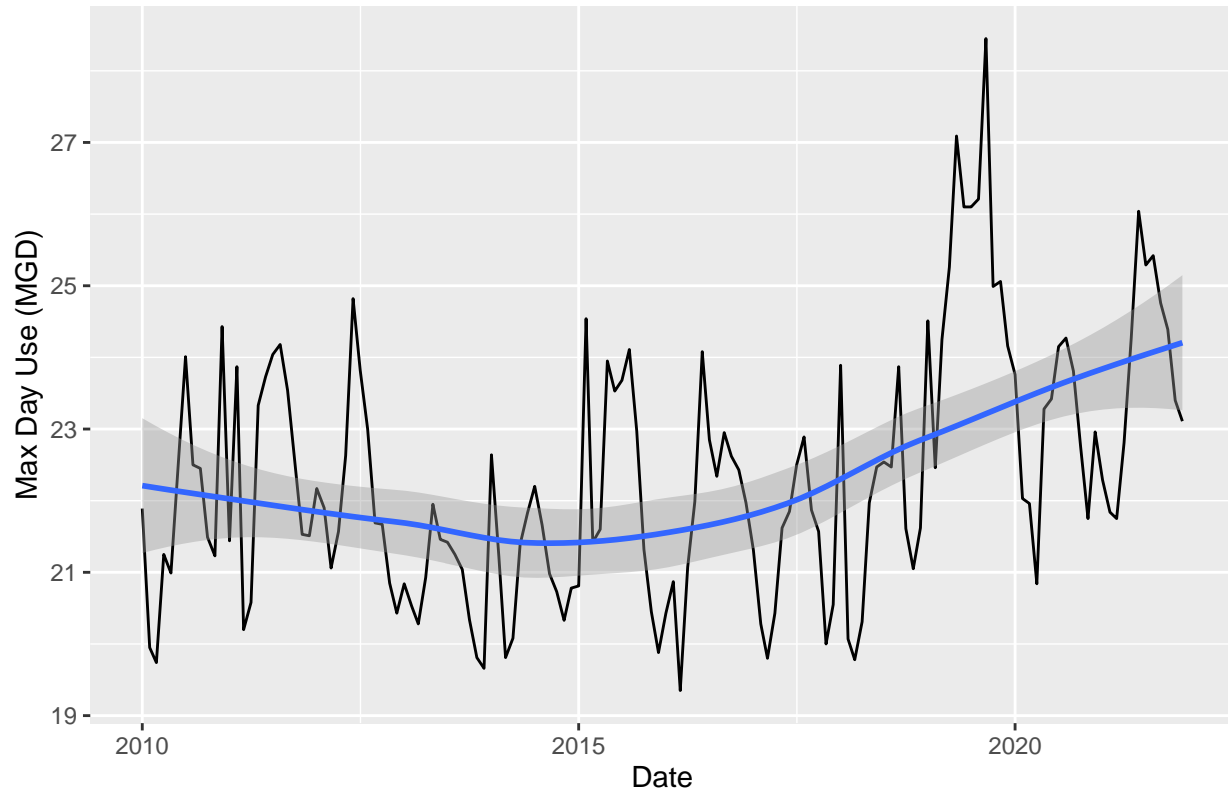
ggplot(water_df_2010_to_2021_v2,aes(y = Max_Day_Use, x=Date)) +
  geom_line()+geom_smooth(method="loess")+
```



```
labs(title="Max Day Use for Asheville from 2010 to 2021", y="Max Day Use (MGD)")
```

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

Max Day Use for Asheville from 2010 to 2021



```
#Manual way of mapping the data (feel free to ignore)
```

```
Asheville_2010 <- scrape.it(2010, '01-11-010')
```

```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2010"
```

```
Asheville_2011 <- scrape.it(2011, '01-11-010')
```

```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2011"
```

```
Asheville_2012 <- scrape.it(2012, '01-11-010')
```

```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2012"
```

```
Asheville_2013 <- scrape.it(2013, '01-11-010')
```

```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2013"
```

```
Asheville_2014 <- scrape.it(2014, '01-11-010')
```

```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2014"
```

```
Asheville_2015 <- scrape.it(2015, '01-11-010')
```

```
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2015"
```

```

Asheville_2016 <- scrape.it(2016, '01-11-010')

## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2016"
Asheville_2017 <- scrape.it(2017, '01-11-010')

## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2017"
Asheville_2018 <- scrape.it(2018, '01-11-010')

## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2018"
Asheville_2019 <- scrape.it(2019, '01-11-010')

## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2019"
Asheville_2020 <- scrape.it(2020, '01-11-010')

## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2020"
Asheville_2021 <- scrape.it(2021, '01-11-010')

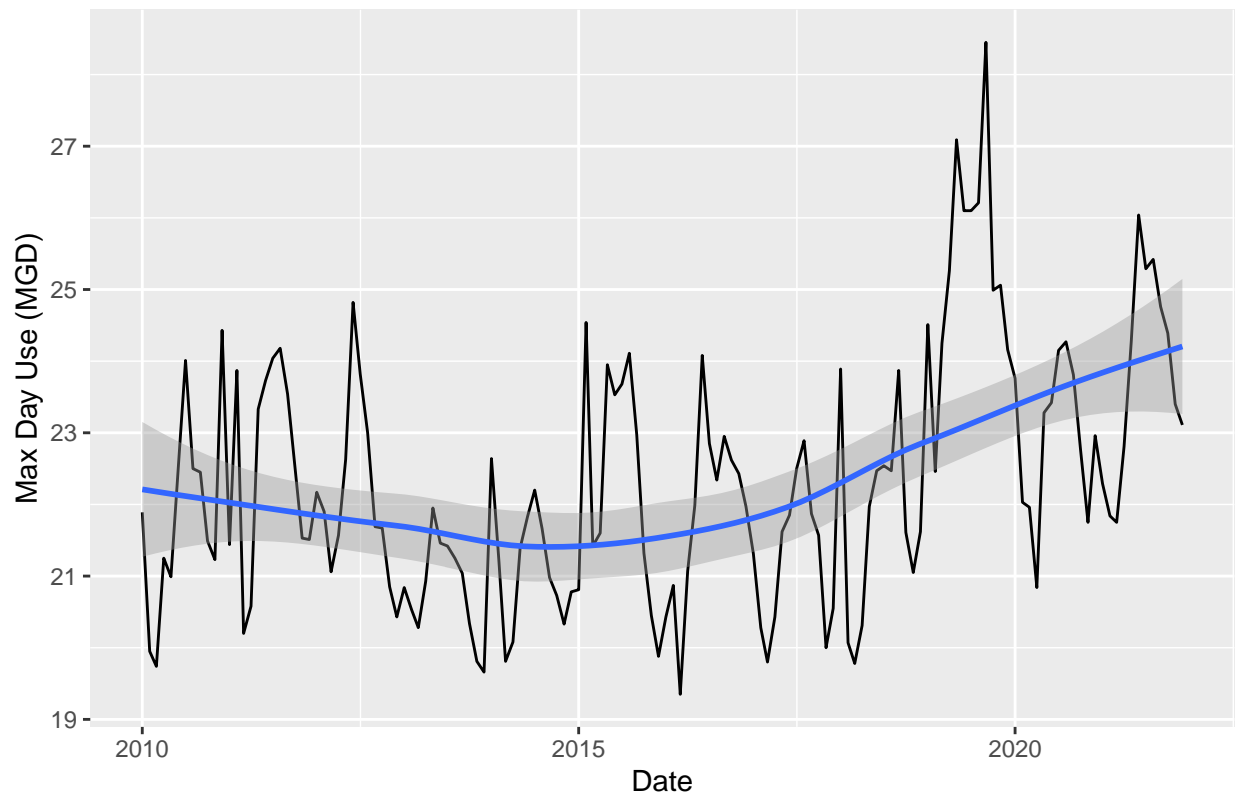
## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=01-11-010&year=2021"
water_df_2010_to_2021_manual<-bind_rows(Asheville_2010, Asheville_2011,Asheville_2012, Asheville_2013,
                                         Asheville_2014, Asheville_2015, Asheville_2016, Asheville_2017,
                                         Asheville_2018,
                                         Asheville_2019, Asheville_2020, Asheville_2021)

#Plot
ggplot(water_df_2010_to_2021_manual,
       aes(y = Max_Day_Use, x=Date)) +
  geom_line()+geom_smooth(method="loess")+
  labs(title="Max Day Use for Asheville from 2010 to 2021", y="Max Day Use (MGD)")

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

```

Max Day Use for Asheville from 2010 to 2021



Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time? > Answer: > Looking at the plot, the water usage in Asheville had a slight decline from 2010 to 2015 then increased from 2015 to 2021.