Assignment 09: Data Scraping

Gaby Antonova

Total points:

OVERVIEW

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

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Fay_09_Data_Scraping.Rmd") prior to submission.

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
getwd()
```

```
## [1] "C:/Users/gabri/OneDrive/Desktop/MPP Coursework/Spring 2022/ENVIRO 872/Environmental_Data_Analyti
```

```
library(tidyverse)
library(rvest)
library(dataRetrieval)
```

```
## Warning: package 'dataRetrieval' was built under R version 4.1.3
```

```
library(lubridate)

mytheme <- theme_classic(base_size = 14) +
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 2019 Municipal Local Water Supply Plan (LWSP):

- Navigate to https://www.ncwater.org/WUDC/app/LWSP/search.php
- Change the date from 2020 to 2019 in the upper right corner.
- 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=2020

Indicate this website as the as the URL to be scraped. (In other words, read the contents into an `rvest` webpage object.)

```
#2

theURL <- read_html('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2020')
```

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:
- Average Daily Use (MGD) - for each month

In the code chunk below scrape these values, assigning them to three 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, with the first value being 36.0100.

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

pswid <- theURL %>%
  html_nodes('td tr:nth-child(1) td:nth-child(5)') %>%
  html_text()

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

max.withdrawals.mgd <- theURL %>% html_nodes('th~ td+ td') %>% html_text()
```

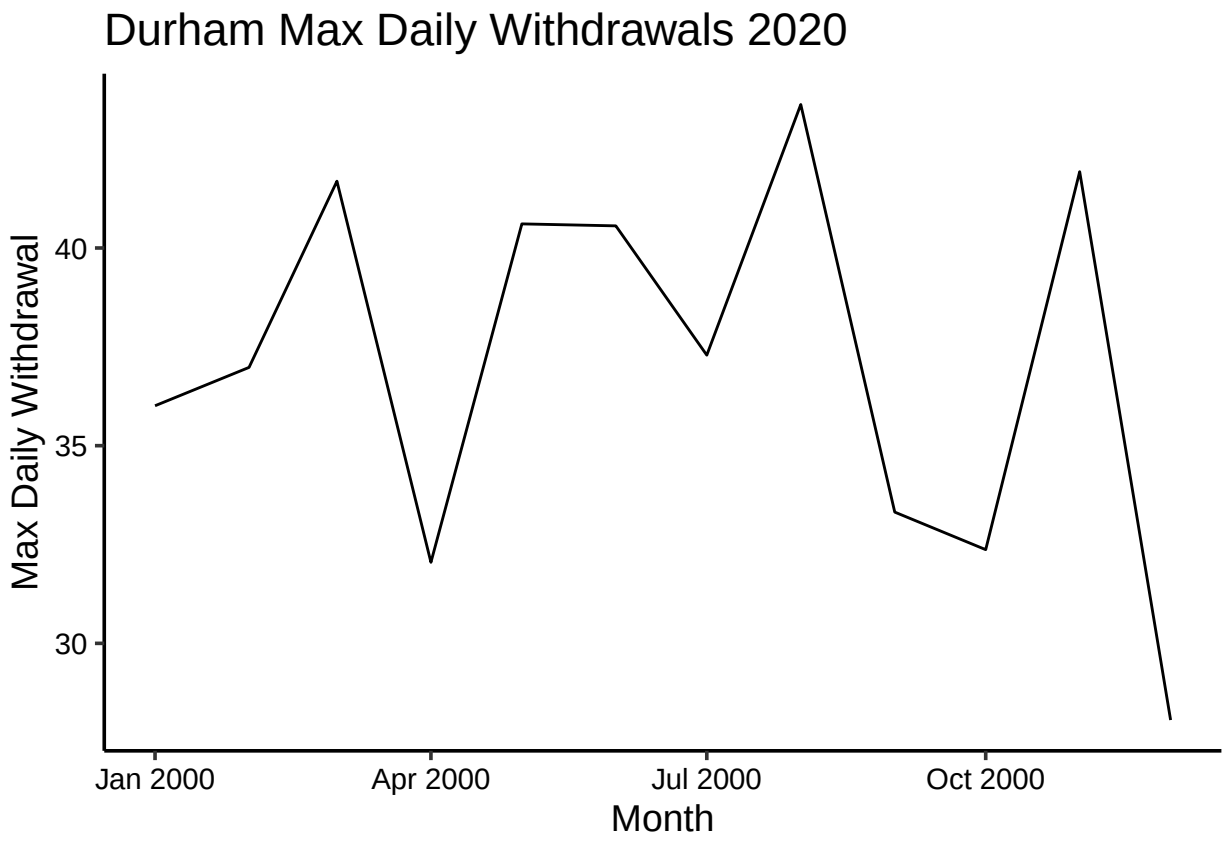
4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

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 order. You can overcome this by creating a month column in the same order the data are scraped: Jan, May, Sept, Feb, etc...

5. Plot the max daily withdrawals across the months for 2020

```
#4

df_electricity <- data.frame("Name" = c(water.system.name),
                             "PSWID" = c(pswid),
                             "Ownership" = c(ownership),
                            "MGD" = as.numeric(max.withdrawals.mgd),
                             "Month" = rep(1:12),
                             "Year" = rep(2000,12))
df_electricity <- mutate(df_electricity, Date = my(paste(Month,"-",Year)))


#5

MaxDailyWithdrawals_plot <- ggplot(df_electricity, aes(Date, MGD) ) +
  geom_line() +
  scale_shape_manual(values = c(1, 12))+
    ylab("Max Daily Withdrawal")+
  xlab("Month")+
  labs(title = "Durham Max Daily Withdrawals 2020")


MaxDailyWithdrawals_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 scraped**.

```
#6.

#create function
scrape.it <- function(the_facility, the_year){


the_scrape_url <- read_html(paste0("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=",
                                   the_facility, '&year=', the_year))
print(the_scrape_url)


#Set the element address variables (determined in the previous step)
water.system.name <- the_scrape_url %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()
pswid <- the_scrape_url %>% html_nodes('td tr:nth-child(1) td:nth-child(5)') %>%
  html_text()
ownership <- the_scrape_url %>%
  html_nodes('div+ table tr:nth-child(2) td:nth-child(4)') %>%
  html_text()
max.withdrawals.mgd <- the_scrape_url %>% html_nodes('th~ td+ td') %>%
```

```
  html_text()



#Convert to a dataframe
df_withdrawals <- data.frame("Month" = c("Jan", "May", "Sep", "Feb", "Jun",
                                         "Oct", "Mar", "Jul", "Nov", "Apr",
                                         "Aug", "Dec"),
                             "Month_Number" = c(1,5,9,2,6,10,3,7,11,4,8,12),
                              "Year" = rep(the_year,12),
                              "Max_Withdrawals_mgd" = as.numeric(max.withdrawals.mgd))

df_withdrawals <- mutate(df_withdrawals, Date = my(paste(Month_Number,"-",Year)))


}
Durham_df <- scrape.it("03-32-010", 2020)
```

```
## {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= ...
```

```
view(Durham_df)
```

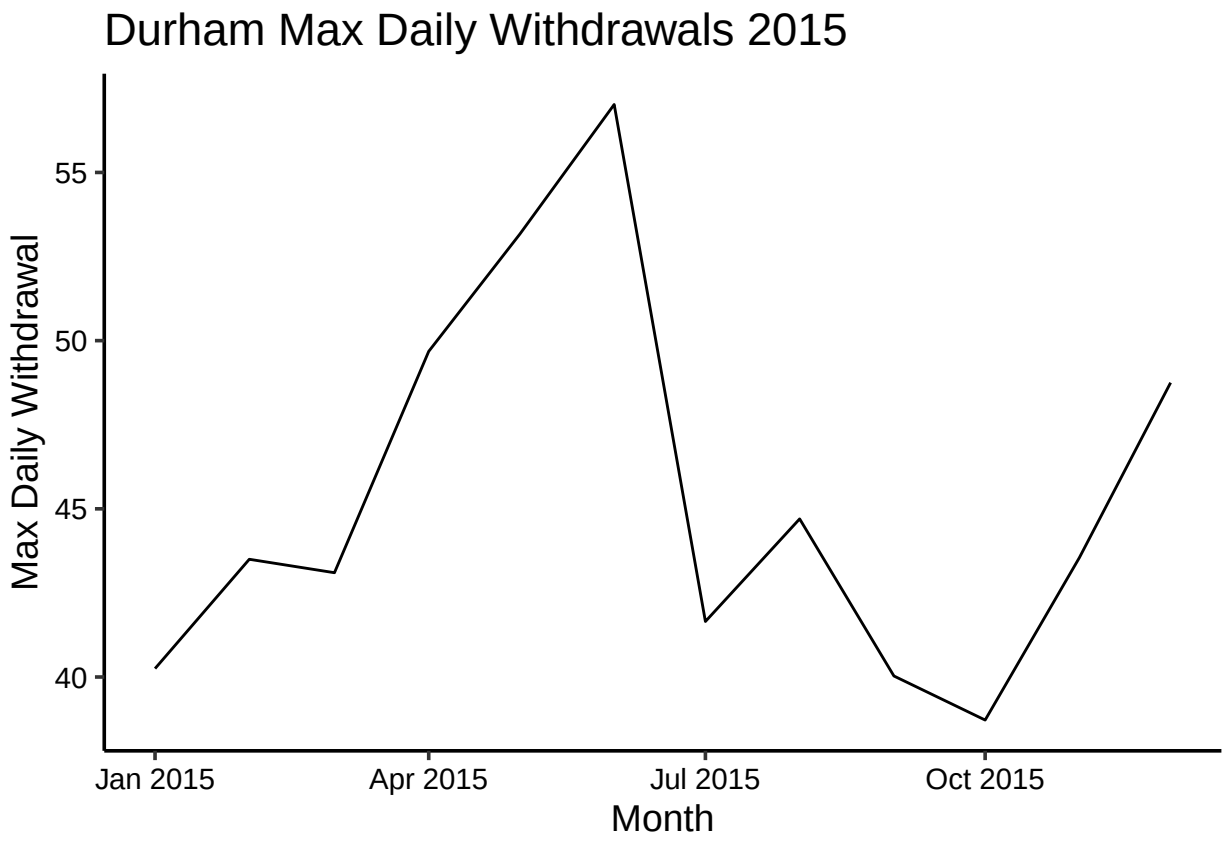
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID=‘03-32-010’) for each month in 2015

```
#7

scrape.it <- function(the_facility, the_year){


the_scrape_url <- read_html(paste0("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=",
                                   the_facility, '&year=', the_year))
print(the_scrape_url)


#Set the element address variables (determined in the previous step)
water.system.name <- the_scrape_url %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()
pswid <- the_scrape_url %>% html_nodes('td tr:nth-child(1) td:nth-child(5)') %>%
  html_text()
ownership <- the_scrape_url %>%
  html_nodes('div+ table tr:nth-child(2) td:nth-child(4)') %>% html_text()
max.withdrawals.mgd <- the_scrape_url %>% html_nodes('th~ td+ td') %>%
  html_text()



#Convert to a dataframe
```

```
df_withdrawals <- data.frame("Month" = c("Jan", "May", "Sep", "Feb", "Jun", "Oct",
                                         "Mar", "Jul", "Nov",
                                         "Apr", "Aug", "Dec"),
                             "Month_Number" = c(1,5,9,2,6,10,3,7,11,4,8,12),
                              "Year" = rep(the_year,12),
                              "Max_Withdrawals_mgd" = as.numeric(max.withdrawals.mgd))

df_withdrawals <- mutate(df_withdrawals, Date = my(paste(Month_Number,"-",Year)))


}
Durham_df <- scrape.it("03-32-010", 2015)
```

```
## {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= ...
```

```
view(Durham_df)

MaxDailyWithdrawals2015_plot <- ggplot(Durham_df, aes(Date, Max_Withdrawals_mgd) ) +
  geom_line() +
  scale_shape_manual(values = c(1, 12))+
    ylab("Max Daily Withdrawal")+
  xlab("Month")+
  labs(title = "Durham Max Daily Withdrawals 2015")


MaxDailyWithdrawals2015_plot
```



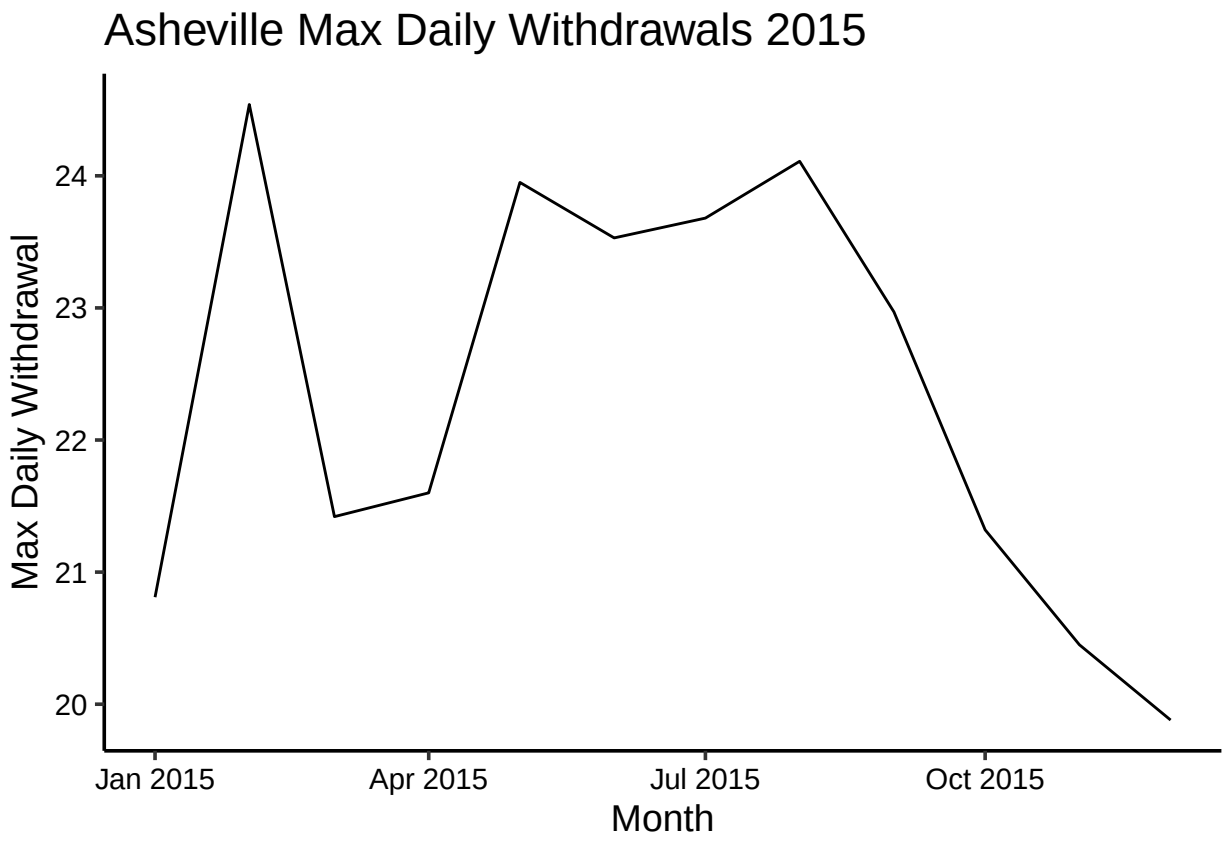
8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares the Asheville to Durham's water withdrawals.

```
#8
scrape.it <- function(the_facility, the_year){


the_scrape_url <- read_html(paste0("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=",
                                   the_facility, '&year=', the_year))
print(the_scrape_url)


#Set the element address variables (determined in the previous step)
water.system.name <- the_scrape_url %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()
pswid <- the_scrape_url %>%
  html_nodes('td tr:nth-child(1) td:nth-child(5)') %>%
  html_text()
ownership <- the_scrape_url %>%
  html_nodes('div+ table tr:nth-child(2) td:nth-child(4)') %>% html_text()
max.withdrawals.mgd <- the_scrape_url %>% html_nodes('th~ td+ td') %>%
  html_text()
```

```
#Convert to a dataframe
df_withdrawals <- data.frame("Month" = c("Jan", "May", "Sep", "Feb", "Jun",
                                         "Oct", "Mar", "Jul", "Nov", "Apr",
                                         "Aug", "Dec"),
                             "Month_Number" = c(1,5,9,2,6,10,3,7,11,4,8,12),
                              "Year" = rep(the_year,12),
                              "Max_Withdrawals_mgd" = as.numeric(max.withdrawals.mgd))

df_withdrawals <- mutate(df_withdrawals,
                         Date = my(paste(Month_Number,"-",Year)))


}
Asheville_df <- scrape.it("01-11-010", 2015)
```

```
## {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= ...
```

```
view(Asheville_df)

AshevilleMaxDailyWithdrawals2015_plot <- ggplot(Asheville_df, aes(Date, Max_Withdrawals_mgd) ) +
  geom_line() +
  scale_shape_manual(values = c(1, 12))+
    ylab("Max Daily Withdrawal")+
  xlab("Month")+
  labs(title = "Asheville Max Daily Withdrawals 2015")


AshevilleMaxDailyWithdrawals2015_plot
```




```
library(cowplot)
```

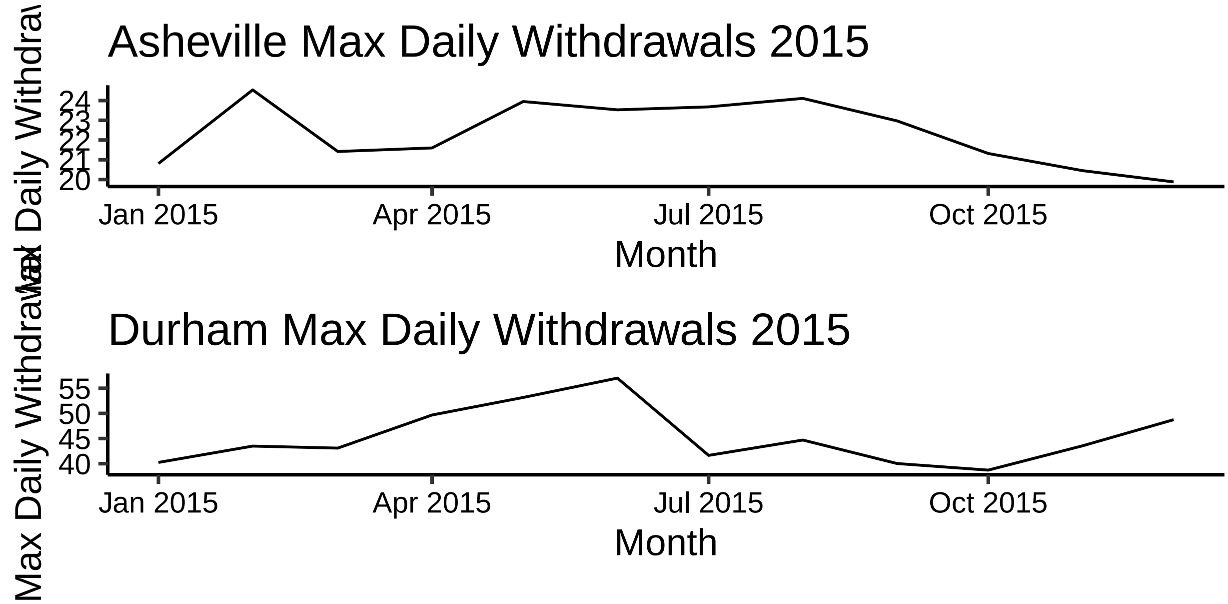
```
##
## Attaching package: 'cowplot'
```

```
## The following object is masked from 'package:lubridate':
##
##     stamp
```

```
library(RColorBrewer)
library(viridis)
```

```
## Loading required package: viridisLite
```

```
combined_plots <-
plot_grid(AshevilleMaxDailyWithdrawals2015_plot, MaxDailyWithdrawals2015_plot,
          nrow = 3, align ="v")

combined_plots
```



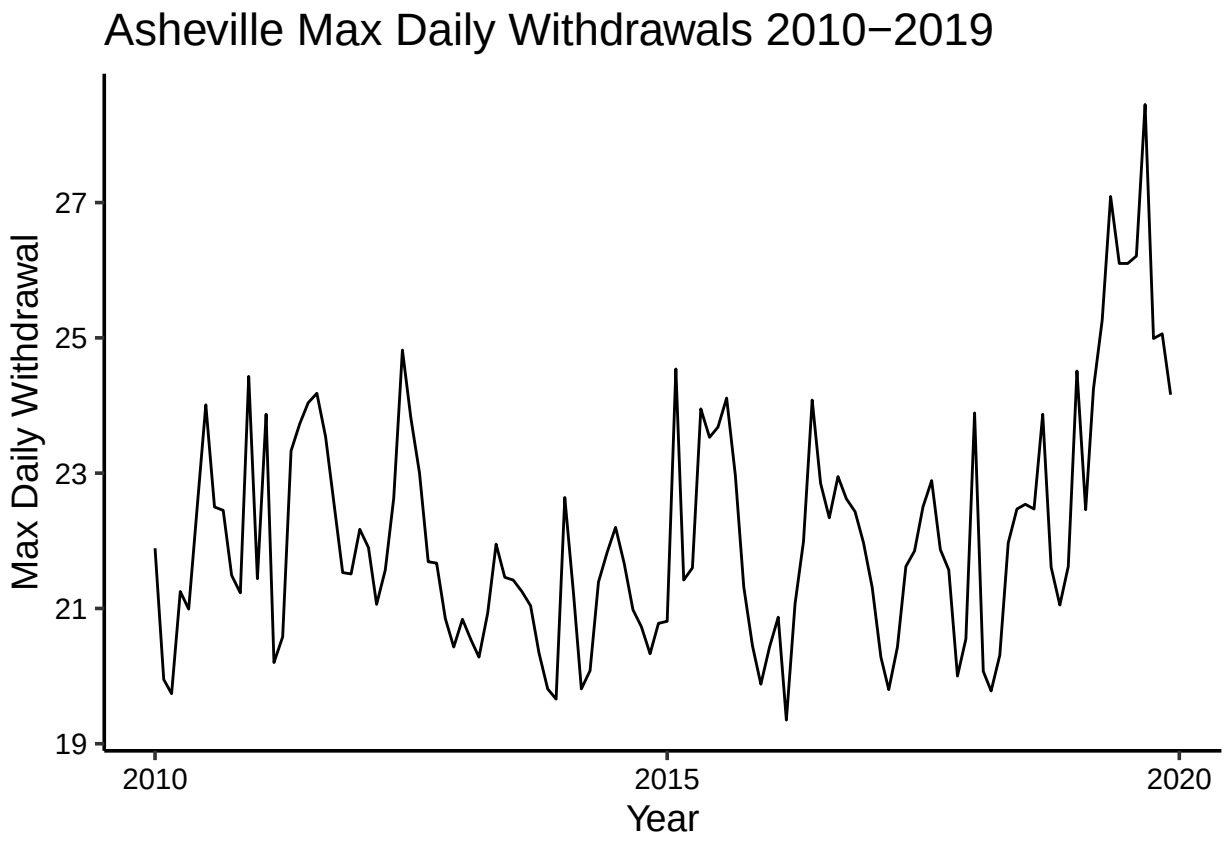
9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

```
#9
the_years = rep(2010:2019)
the_site = "01-11-010"

the_dfs <- map(the_years, scrape.it, the_facility=the_site)
```

```
## {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= ...
## {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= ...
## {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= ...
## {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= ...
```

```
## {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= ...
## {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= ...
## {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= ...
## {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= ...
## {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= ...
## {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= ...
```

```r
the_df <- bind_rows(the_dfs)

ggplot(the_df, aes(Date, Max_Withdrawals_mgd) ) +
  geom_line() +
  scale_shape_manual(values = c(1, 12))+
    ylab("Max Daily Withdrawal")+
  xlab("Year")+
  labs(title = "Asheville Max Daily Withdrawals 2010-2019")
```



Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time?

There appears to be an increasing trend in water usage in the last few years (2017-2020), however the water usage trend appeared to be relatively flat between 2010 and 2017.