Assignment 10: Data Scraping

Nancy Bao

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_10_Data_Scraping.Rmd") prior to submission.

The completed exercise is due on Tuesday, April 6 at 11:59 pm.

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
#Check working directory
getwd()
```

```
## [1] "/Users/Nancy/Desktop/Semester 4/ENV 872L/Environmental_Data_Analytics_2021/Assignments"
```

```
#Load packages
library(tidyverse)
library(rvest)
library(lubridate)
library(viridis)
library(dataRetrieval)
#Set ggplot theme
A10theme<- theme_linedraw(base_size=12,base_family="")+
           theme(plot.title=element_text(size=14,
                                         face="bold",
                                         color="black",
                                         hjust=0.5),
                 axis.text = element_text(color = "black"),
                 axis.title = element_text(color= "black",face= "bold"),
                 legend.position = "top")
```

```
#Set the theme
theme_set(A10theme)
```

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=2019

Indicate this website as the URL to be scraped.

```
#2 Scrape data from the NC DEQ's local water supply planning website
water.website <-
  read_html('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2019')
water.website
```

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

```
#I called the scraped website: water.website
```

3. The data we want to collect are listed below:

- From the "System Information" section:

- Water system name

- PSWID

- Ownership

- From the "Water Supply Sources" section:

- Maximum monthly withdrawals (MGD)

In the code chunk below scrape these values into the supplied variable names.

```
#3
#System Information section:
water_system <-water.website %>%
              html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>%
              html_text()
water_system
```

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

```
#I called the water system name: water_system
PSWID <- water.website %>%
              html_nodes('td tr:nth-child(1) td:nth-child(5)') %>%
              html_text()
PSWID
```

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

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

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

```
#Water Supply Sources section:
#I took the values from the Max Day Use (MGD) columns.
max.monthly.withdrawal.MGD <- water.website %>%
            html_nodes('th~ td+ td') %>%
            html_text()
max.monthly.withdrawal.MGD
```

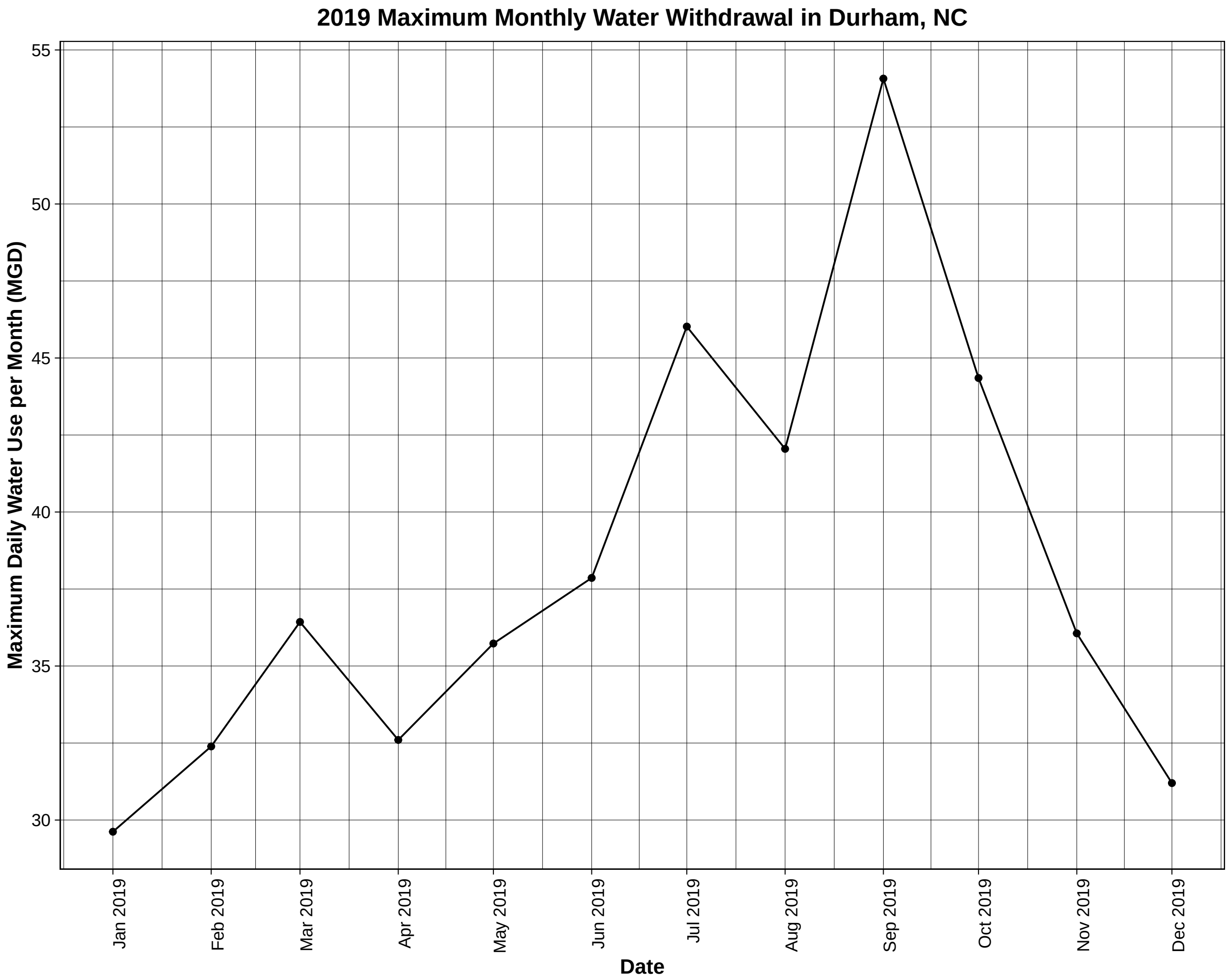
```
##  [1] "29.6200" "35.7300" "54.0700" "32.3900" "37.8600" "44.3500" "36.4300"
##  [8] "46.0200" "36.0600" "32.6000" "42.0500" "31.2000"
```

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.)

   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 2019.

```
#4 Converting scraped data to dataframe
Durham_h2o_df <-data.frame("Water_System"= water_system,
                           "PSW_ID"= PSWID,
                           "Ownership"= ownership,
                           "Max_Monthly_Withdrawals_mgd"=
                             as.numeric(max.monthly.withdrawal.MGD)) %>%
                  mutate(Month = c(1,5,9,2,6,10,3,7,11,4,8,12),#added month column manually
                           Year = 2019) %>%
                  mutate(Date =
                        my(paste0(Month,"-",Year))) %>%
                        arrange(Date) #arranged by ascending date
#5 Max Daily Withdrawals across the months for 2019
monthly.max.withdrawals.plot <- ggplot(Durham_h2o_df,
                                       aes(x=Date,y=Max_Monthly_Withdrawals_mgd))+
                                 geom_point(alpha=1,size=1.5)+
                                 geom_line()+
                                 labs(y="Maximum Daily Water Use per Month (MGD)", x="Date",
                                 title="2019 Maximum Monthly Water Withdrawal in Durham, NC")+
  scale_x_date(date_breaks = "1 month",date_labels = "%b %Y")+
  theme(axis.text.x=element_text(angle=90, hjust=1))
#adjusted text angle, so all text would be visible and not crowded together
##x-scale is by 1 month and I changed the format to 3-letter month and 4-digit year
plot(monthly.max.withdrawals.plot)
```



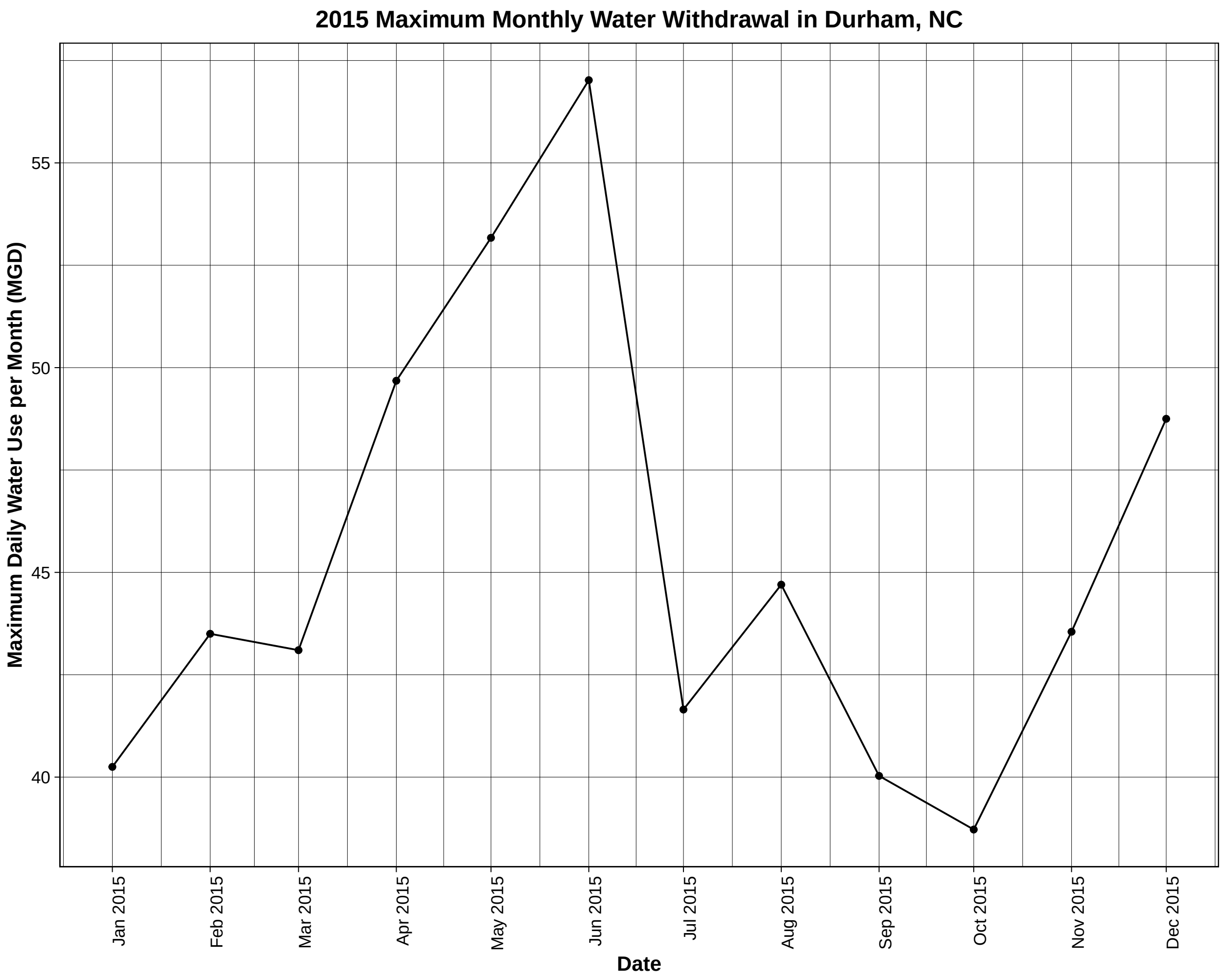
6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. Be sure to modify the code to reflect the year and data scraped.

```
#6.Scraping Function
scrape.pws <-function(PSW_ID,Year_){
                water.website1<-read_html(
                  paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=',
                         PSW_ID,'&year=',Year_)) #function scrapes for PSW_ID and the designated year
    #Scraped data from website
    water_system <-water.website1 %>%
              html_nodes('div+ table tr:nth-child(1) td:nth-child(2)') %>%
              html_text()
    PSWID <- water.website1 %>%
              html_nodes('td tr:nth-child(1) td:nth-child(5)') %>%
              html_text()
    ownership <- water.website1 %>%
              html_nodes('div+ table tr:nth-child(2) td:nth-child(4)') %>%
              html_text()
    max.monthly.withdrawal.MGD <- water.website1 %>%
            html_nodes('th~ td+ td') %>%
            html_text()
#Convert to dataframe
```

```
    h20_df<-data.frame("Water_System"= water_system,
                           "PSW_ID"= PSWID,
                           "Ownership"= ownership,
                           "Max_Monthly_Withdrawals_mgd"=
                            as.numeric(max.monthly.withdrawal.MGD)) %>%
                            mutate(Month = c(1,5,9,2,6,10,3,7,11,4,8,12),
                                   #added month column manually
                                   Year = !!Year_)%>%
                  mutate(Date =
                        my(paste0(Month,"-",Year))) %>%
                        arrange(Date) #arranged by ascending date
     return(h20_df)
}
```

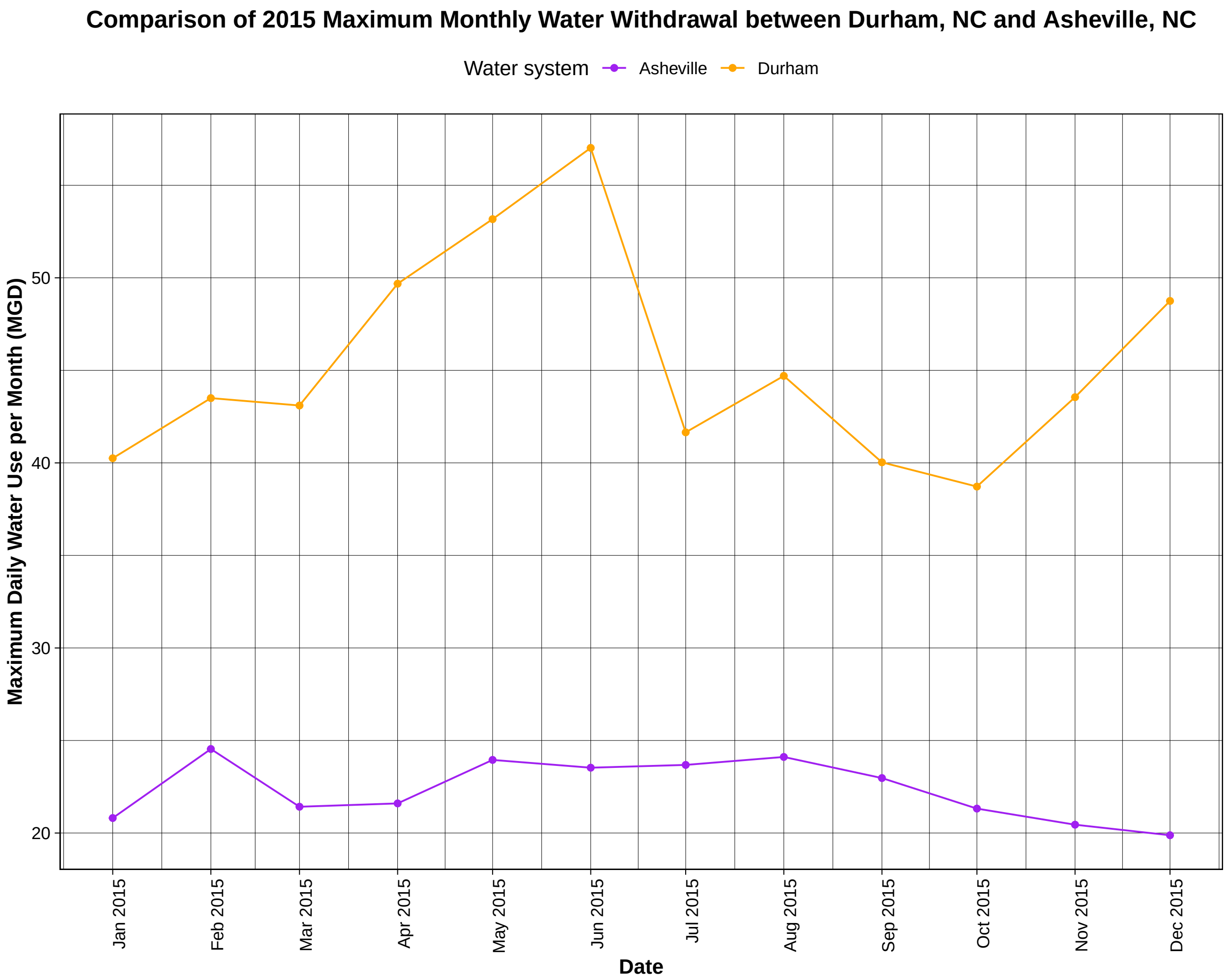
7. Use the function above to extract and plot max daily withdrawals for Durham for each month in 2015

```
#7
#Extract max daily withdrawals for Durham for each month in 2015
#I used the function I created above by scrape.pws('PSW_ID',Year)
#I filled in the PWS id and 2015 to the scrape.pws() function
Durham2015_h2o_df<-scrape.pws('03-32-010',2015)
#Plot max daily withdrawals for each month in 2015 for Durham, NC
monthly.2015.max.withdrawals.plot <- ggplot(Durham2015_h2o_df,
                                       aes(x=Date,y=Max_Monthly_Withdrawals_mgd))+
                                 geom_point(alpha=1,size=1.5)+
                                 geom_line()+
                                 labs(y="Maximum Daily Water Use per Month (MGD)", x="Date",
                                 title="2015 Maximum Monthly Water Withdrawal in Durham, NC")+
  scale_x_date(date_breaks = "1 month",date_labels = "%b %Y")+
  theme(axis.text.x=element_text(angle=90, hjust=1))
#I broke the x scale down by 1 month and changed the format to 3-letter month and 4-digit year
plot(monthly.2015.max.withdrawals.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.

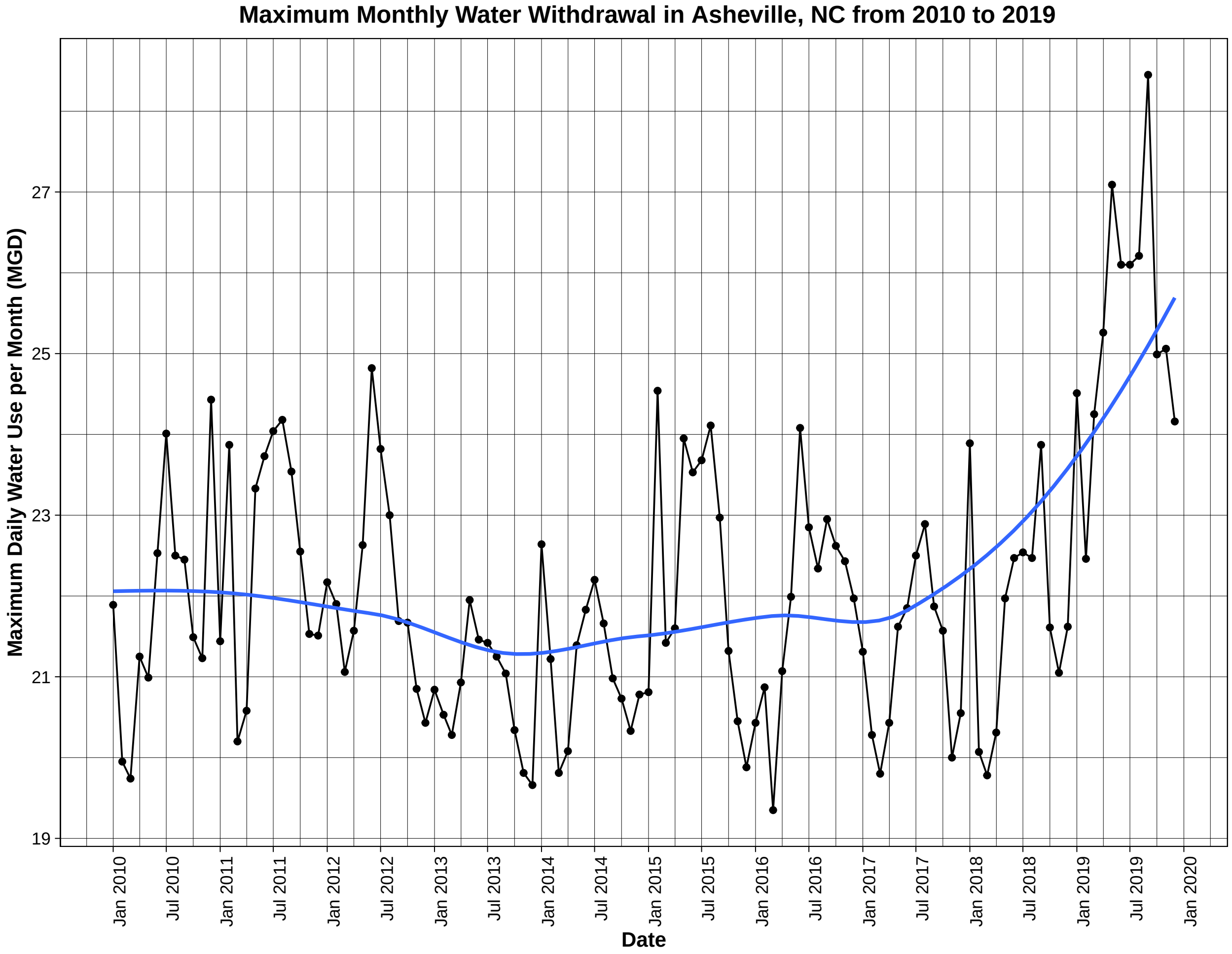
```
#8
#Use function extract data for Asheville (PWSID=01-11-01) in 2015
Asheville2015_h2O_df <-scrape.pws('01-11-010',2015)
#Combine data with Durham data using rbind
NC_h2O_df <-rbind(Durham2015_h2o_df,Asheville2015_h2O_df)
#Plot combined dataframe: NC_h2O_df
combined.monthly.2015.max.withdrawals.plot <- ggplot(NC_h2O_df,
                                       aes(x=Date,y=Max_Monthly_Withdrawals_mgd,color=Water_System))+
                                 geom_point(alpha=1,size=1.5)+
                                 geom_line()+
                                 labs(y="Maximum Daily Water Use per Month (MGD)", x="Date",
                           title="Comparison of 2015 Maximum Monthly Water Withdrawal between Durham, N
  scale_x_date(date_breaks = "1 month",date_labels = "%b %Y")+
  scale_color_manual(values = c("purple","orange"))+
  theme(axis.text.x=element_text(angle=90, hjust=1))
#I changed the color of the plots to purple and orange for more contrast
plot(combined.monthly.2015.max.withdrawals.plot)
```



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
#Setting inputs and used map(rep()) for multiple year interval
pwsid <-'01-11-010'
Asheville_10_19_h2O_df <-map(rep(2010:2019),scrape.pws,PSW_ID=pwsid) %>% bind_rows()
#
Asheville_2010_2019_plot<-ggplot(Asheville_10_19_h2O_df,
                                 aes(x=Date,y=Max_Monthly_Withdrawals_mgd))+
                                 geom_point(alpha=1,size=1.5)+
                                 geom_line()+
                                 geom_smooth(method="loess",se=FALSE)+
                                 labs(y="Maximum Daily Water Use per Month (MGD)",
                                      x="Date",
                         title=" Maximum Monthly Water Withdrawal in Asheville, NC from 2010 to 2019",
                                 caption="Blue trend line")+
                                 scale_x_date(date_breaks = "6 months",date_labels = "%b %Y")+
                                 theme(axis.text.x=element_text(angle=90, hjust=1))
plot(Asheville_2010_2019_plot)
```

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



Blue trend line

```
#I decided to add a geom_line to show the trends between each point and
#then I added an overall geomsmooth()
#to better show the trend of all the data points
```

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

Yes, Asheville, NC does have a trend in water usage over time. From the smoothed line in the figure above, there appears to be a decrease in water usage from 2010 to 2013 and an increase in water usage over time starting from around early 2014 and increasing more rapidly in water usage (steeper smoothed line) from 2016 to 2019. Within each year, the summer months from June to August peak in water use compared to the winter months, January to March.