Assignment 10: Data Scraping

Karly Nocera

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
getwd()
```

```
## [1] "Z:/Environmental_Data_Analytics_2021"
```

```
# load
library(tidyverse)
library(lubridate)
library(rvest)

# Set theme
mytheme <- theme_classic() +
  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=2019

Indicate this website as the as the URL to be scraped.

```
#2

webpage <- read_html('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2019')
webpage
```

```
## {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 "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

water_system <- webpage %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()

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

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

max_month_wdrawl <- webpage %>%
  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.)

   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...

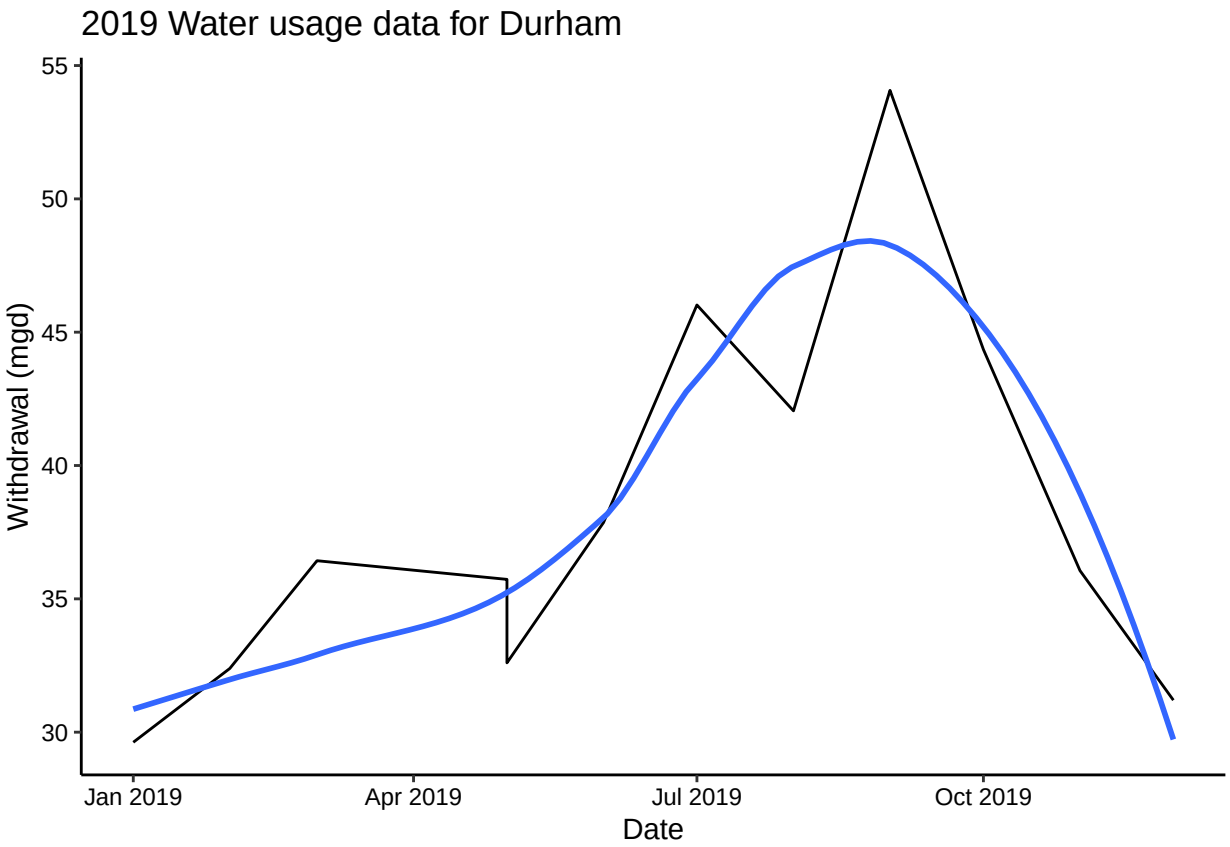
5. Plot the max daily withdrawals across the months for 2019.

```
#4

#identify order of month withdrawals
max_month_wdrawl
```

```
##  [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"
```

```
## jan, may, sep, feb, jun, oct, mar, jul, nov, apr, aug, dec

#Create a dataframe of withdrawals
df_withdrawals <- data.frame("Month" = c(1, 5, 9, 2, 6, 10, 3, 7, 11, 5, 8, 12), #corresponding months
                             "Year" = rep(2019,12),     #make 12 instances of 2019
                             "Max_Monthly_Withdrawal" = as.numeric(max_month_wdrawl))

#Modify the dataframe to include the ownership and water system names as well as the date (as date obje
df_withdrawals <- df_withdrawals %>%
  mutate(Owndership = !!owner,   #don't set it as a string, it refers as a variable
         Water_System = !!water_system,
         Date = my(paste(Month,"-",Year)))

#5

ggplot(df_withdrawals,aes(x=Date,y=Max_Monthly_Withdrawal)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2019 Water usage data for",water_system),
       y="Withdrawal (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 data scraped.

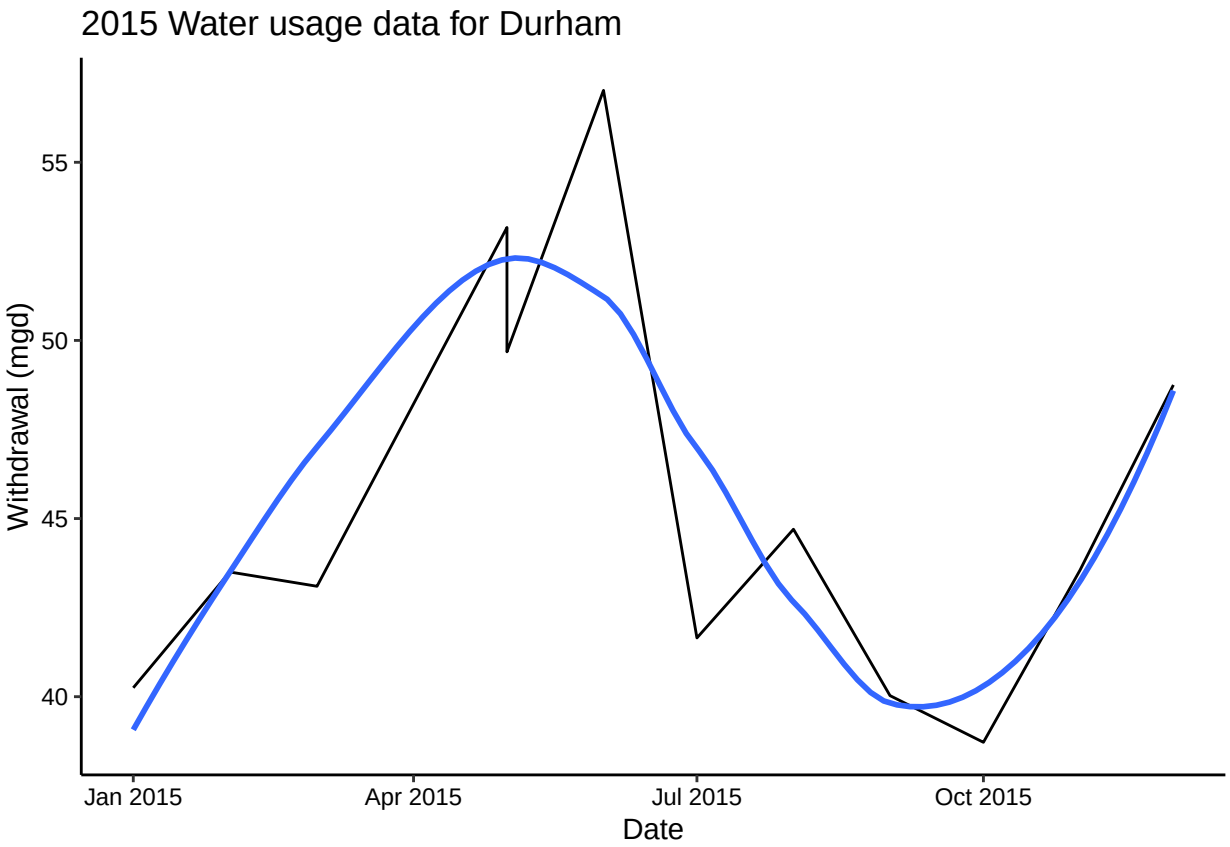
```
#6.

scrape <- function(year, PSWID){

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

  #Set the element address variables (determined in the previous step)
  water_system_tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'
  pswid_tag <- 'td tr:nth-child(1) td:nth-child(5)'
  owner_tag <- 'div+ table tr:nth-child(2) td:nth-child(4)'
  max_month_wdrawl_tag <- 'th~ td+ td'

  #Scrape the data items
  water_system <- website %>% html_nodes(water_system_tag) %>% html_text()
  PSWID <- website %>%   html_nodes(pswid_tag) %>%  html_text()
  owner <- website %>% html_nodes(owner_tag) %>% html_text()
  max_month_wdrawl <- website %>% html_nodes(max_month_wdrawl_tag) %>% html_text()

  #Convert to a dataframe
  df_withdrawals <- data.frame("Month" = c(1, 5, 9, 2, 6, 10, 3, 7, 11, 5, 8, 12),
                               "Year" = rep(year,12),
```

```
                              "Max_Monthly_Withdrawal" = as.numeric(max_month_wdrawl)) %>%
    mutate(Owndership = !!owner,
           Water_System = !!water_system,
           Date = my(paste(Month,"-",Year)))


  #scraping etiquette
  Sys.sleep(5)

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

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

```
#7

durham_df_2015 <- scrape(2015, '03-32-010')
view(durham_df_2015)

ggplot(durham_df_2015,aes(x=Date,y=Max_Monthly_Withdrawal)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2015 Water usage data for",water_system),
       y="Withdrawal (mgd)",
       x="Date")
```

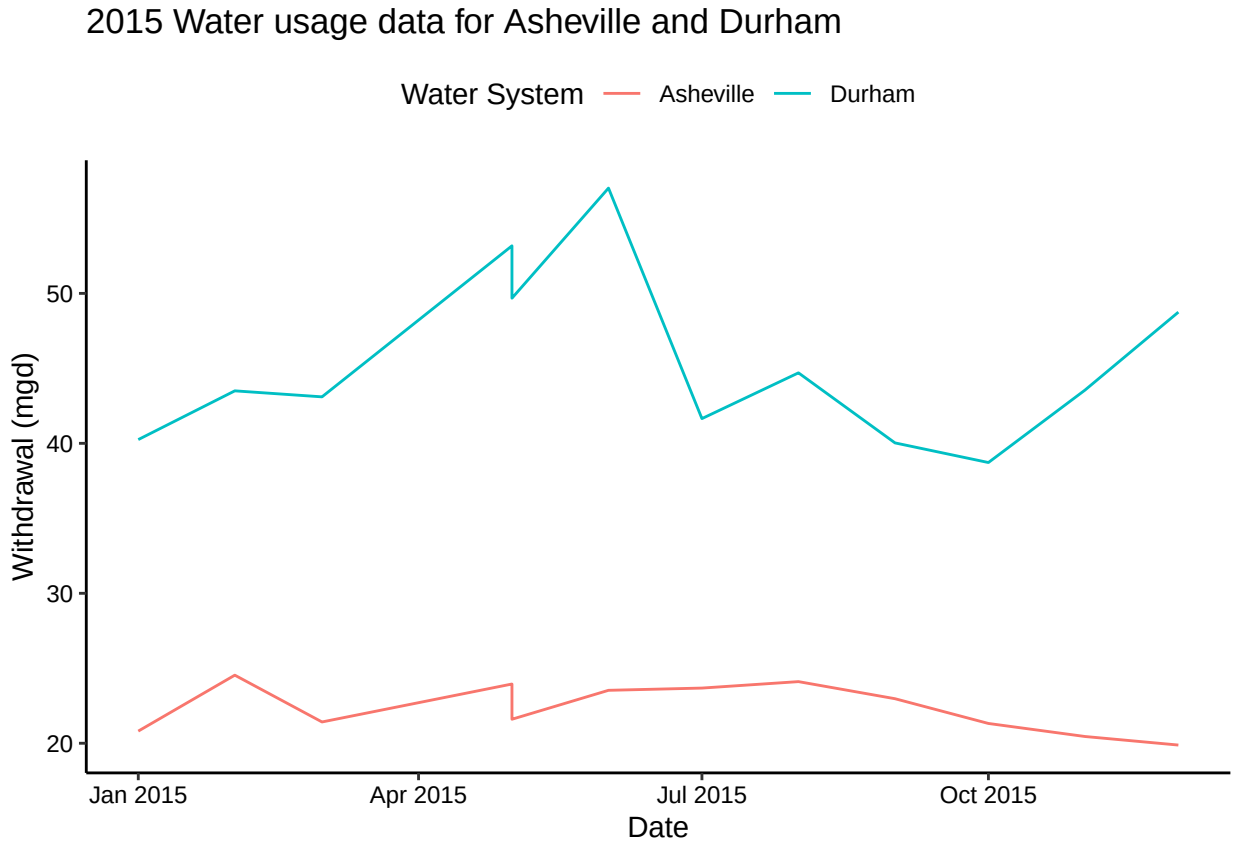
```
## `geom_smooth()` using formula 'y ~ x'
```



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 the Asheville to Durham's water withdrawals.

```
#8

#extract Asheville
ash_df_2015 <- scrape(2015, '01-11-010')
view(ash_df_2015)

#combine with Durham
combine <- rbind(durham_df_2015, ash_df_2015)

#plot comparison of Asheville to Durham withdrawals

ggplot(combine,aes(x=Date,y=Max_Monthly_Withdrawal, color = Water_System)) +
  geom_line(aes(fill = Water_System)) +
  labs(title = paste("2015 Water usage data for Asheville and Durham"),
       y="Withdrawal (mgd)",
       x="Date",
       color = "Water System")
```

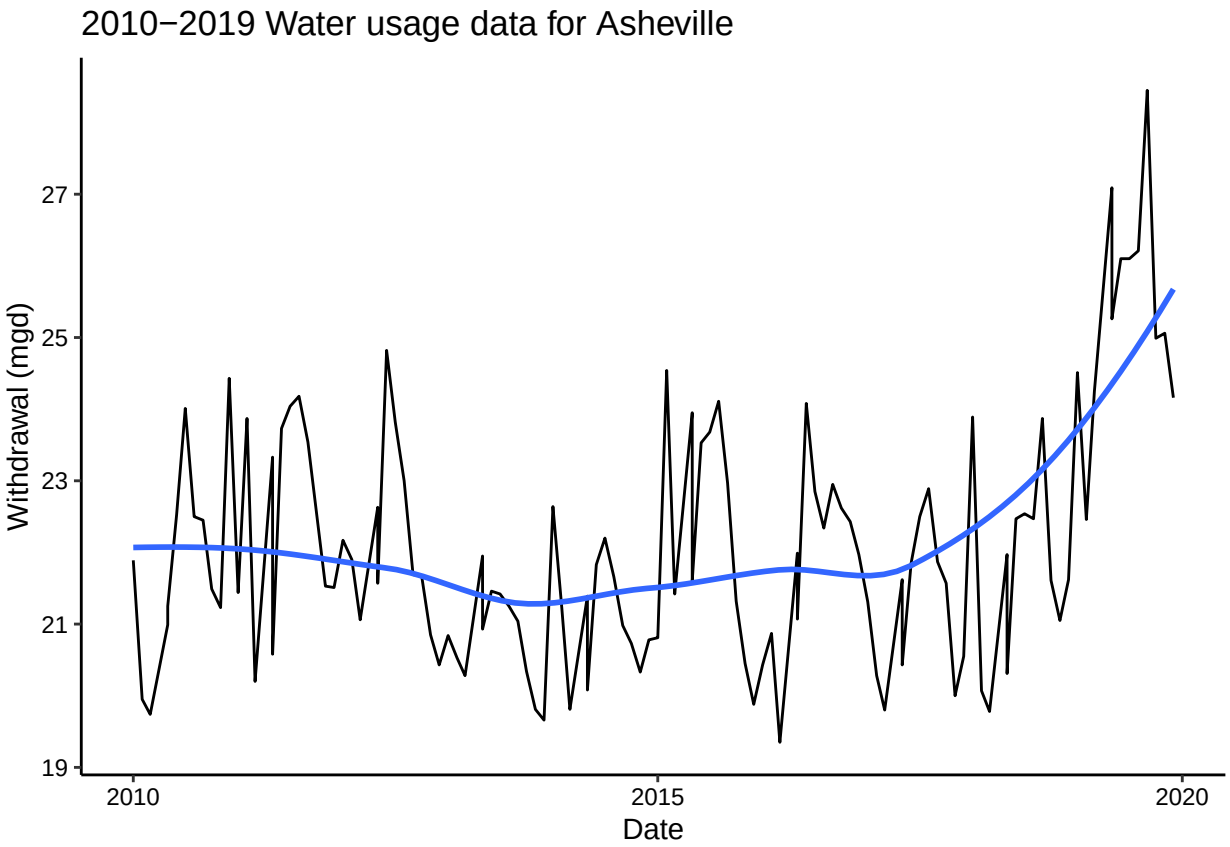
```
## Warning: Ignoring unknown aesthetics: fill
```



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

#create years wanted
years <- rep(2010:2019)

#pwsid for Asheville
pswid_ash <- '01-11-010'

#use purrr's map function
the_dfs <- map(years,scrape,PSWID=pswid_ash)

#Conflate the returned dataframes into a single dataframe
ash_2010.2019_df <- bind_rows(the_dfs)


#Plot
ggplot(ash_2010.2019_df,aes(x=Date,y=Max_Monthly_Withdrawal)) +
  geom_line() +
  geom_smooth(method="loess",se=FALSE) +
  labs(title = paste("2010-2019 Water usage data for Asheville"),
       y="Withdrawal (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, it appears that the plot indicates an increasing trend of Asheville's water usage over time.