Assignment 09: Data Scraping

Curtis Cha

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
library(pacman)
```

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

```
pacman::p_load(tidyverse, rvest, lubridate, ggplot2, cowplot)
```

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

website <- 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 <- 'div+ table tr:nth-child(1) td:nth-child(2)'
pswid <- "td tr:nth-child(1) td:nth-child(5)"
ownership <- "div+ table tr:nth-child(2) td:nth-child(4)"
max.withdrawals.mgd <- "th~ td+ td"
```

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...

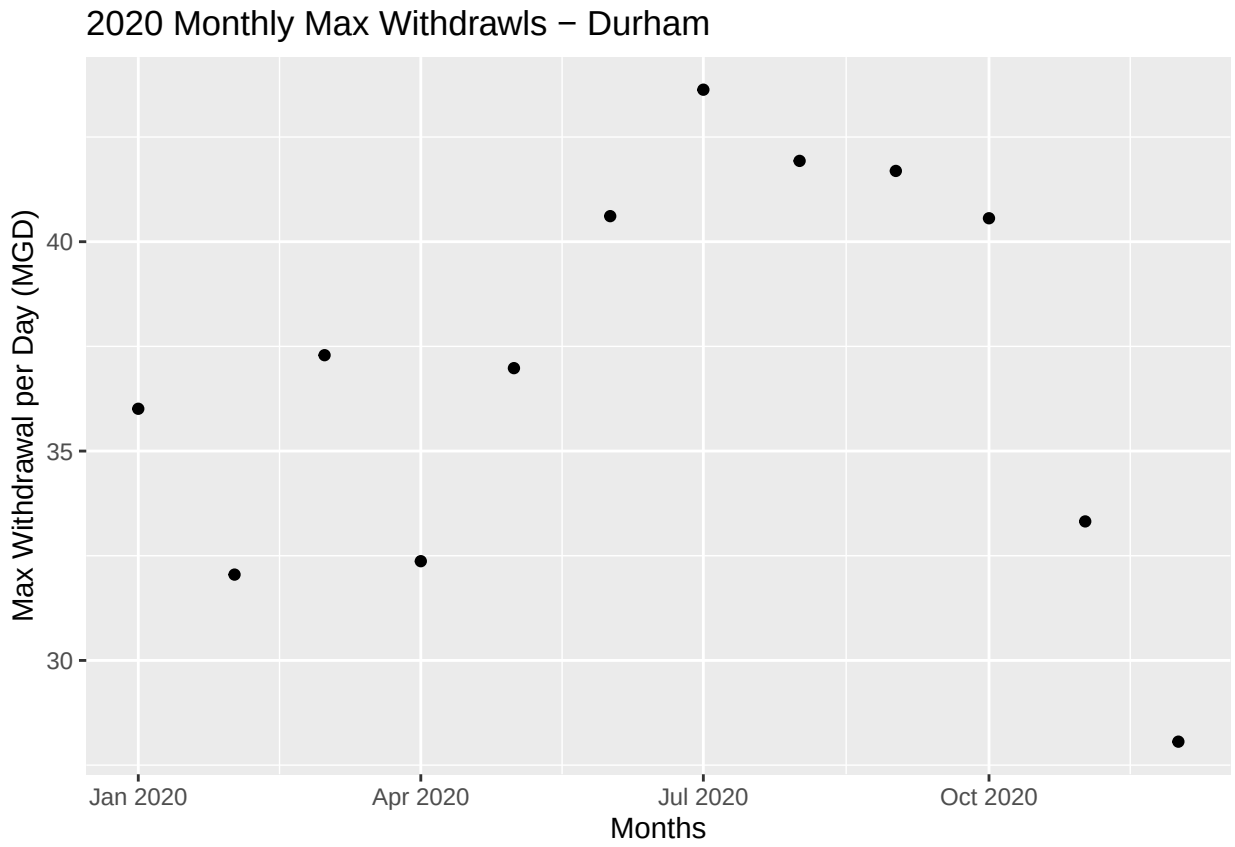
5. Plot the max daily withdrawals across the months for 2020

```
#4

water_system <- website %>%
  html_nodes(water.system.name) %>% html_text()
ID <- website %>%
    html_nodes(pswid) %>%  html_text()
Ownership <- website %>%
   html_nodes(ownership) %>% html_text()
max_withdrawals <- website %>%
    html_nodes(max.withdrawals.mgd) %>% html_text()

months <- c("Jan", "May", "Sep", "Feb", "Jun",
            "Oct", "Mar", "Jul", "Nov", "Apr", "Aug", "Dec")

df <- data.frame("Water System" = rep(water_system, 12),
                 "PSWID" = rep(ID,12),
                 "Ownership" = rep(Ownership,12),
                 "Month" = my(paste(months,"-2020", sep ='')),
                 "Max_Withdrawals_mgd" = as.numeric(max_withdrawals))

df
```

```
##    Water.System     PSWID    Ownership      Month Max_Withdrawals_mgd
## 1        Durham 03-32-010 Municipality 2020-01-01               36.01
## 2        Durham 03-32-010 Municipality 2020-05-01               36.98
```

```
## 3        Durham 03-32-010 Municipality 2020-09-01              41.69
## 4        Durham 03-32-010 Municipality 2020-02-01              32.05
## 5        Durham 03-32-010 Municipality 2020-06-01              40.61
## 6        Durham 03-32-010 Municipality 2020-10-01              40.56
## 7        Durham 03-32-010 Municipality 2020-03-01              37.29
## 8        Durham 03-32-010 Municipality 2020-07-01              43.63
## 9        Durham 03-32-010 Municipality 2020-11-01              33.32
## 10       Durham 03-32-010 Municipality 2020-04-01              32.37
## 11       Durham 03-32-010 Municipality 2020-08-01              41.93
## 12       Durham 03-32-010 Municipality 2020-12-01              28.06
```

```r
#5

durham_2020 <- ggplot(df, aes(x = Month, y = Max_Withdrawals_mgd)) +
  geom_point() +
  labs(title = "2020 Monthly Max Withdrawls - Durham",
       x = "Months",
       y = "Max Withdrawal per Day (MGD)")

durham_2020
```



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**.

```r
#6.

scrape.it <- function(pwsid, year) {
```

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

  #html_elemnts
  water.system.name <- 'div+ table tr:nth-child(1) td:nth-child(2)'
  pswid <- "td tr:nth-child(1) td:nth-child(5)"
  ownership <- "div+ table tr:nth-child(2) td:nth-child(4)"
  max.withdrawals.mgd <- "th~ td+ td"

  water_system <- website %>% html_nodes(water.system.name) %>% html_text()
  ID <- website %>% html_nodes(pswid) %>%  html_text()
  Ownership <- website %>% html_nodes(ownership) %>% html_text()
  max_withdrawals <- website %>% html_nodes(max.withdrawals.mgd) %>% html_text()

  #dataframe
  months <- c("Jan", "May", "Sep", "Feb", "Jun",
              "Oct", "Mar", "Jul", "Nov", "Apr", "Aug", "Dec")

  df <- data.frame("Water System" = rep(water_system, 12),
                   "PSWID" = rep(ID,12),
                   "Ownership" = rep(Ownership,12),
                   "Month" = my(paste(months,"-", year, sep ='')),
                   "Max_Withdrawals_mgd" = as.numeric(max_withdrawals))

  return(df)
}
```

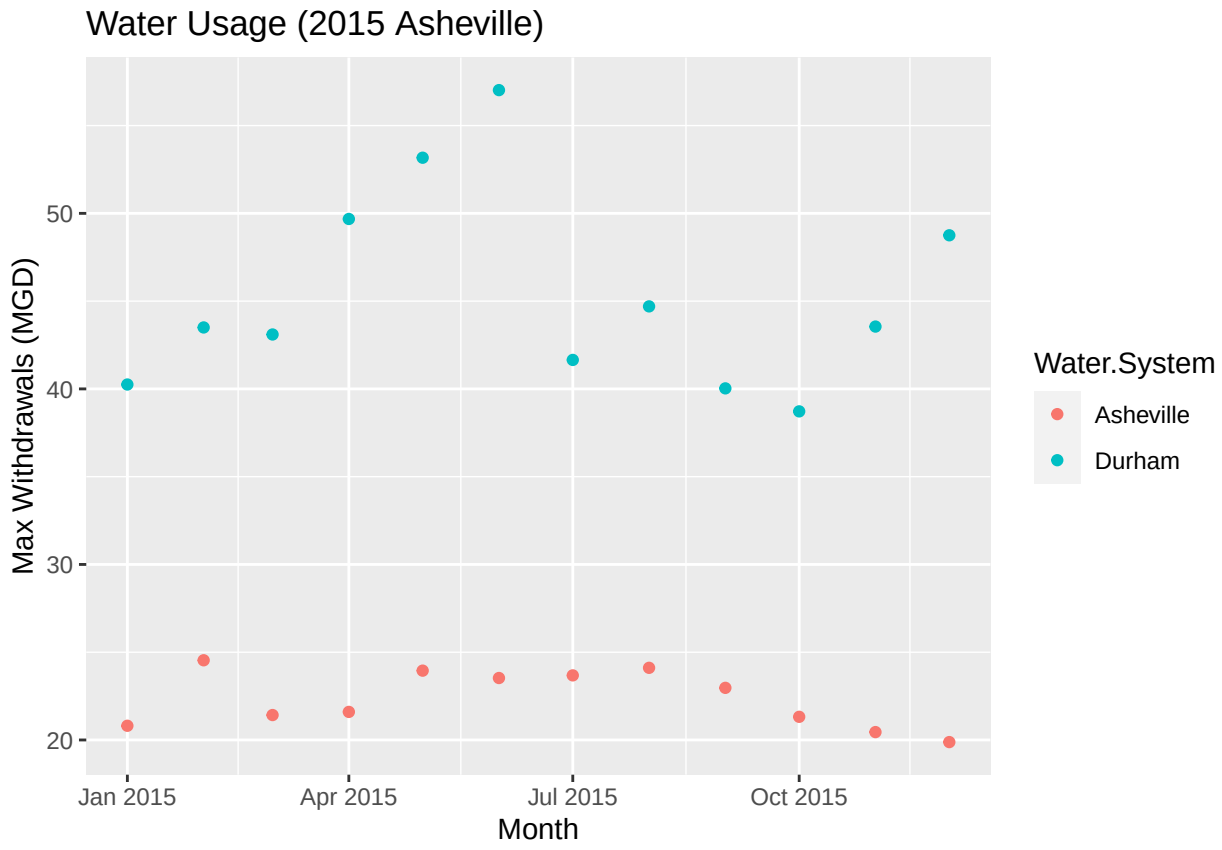
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID=‘03-32-010’) for each month in 2015

```
#7
durham_2015 <- scrape.it('03-32-010', '2015')
durham_2015
```

```
##    Water.System     PSWID    Ownership      Month Max_Withdrawals_mgd
## 1        Durham 03-32-010 Municipality 2015-01-01               40.25
## 2        Durham 03-32-010 Municipality 2015-05-01               53.17
## 3        Durham 03-32-010 Municipality 2015-09-01               40.03
## 4        Durham 03-32-010 Municipality 2015-02-01               43.50
## 5        Durham 03-32-010 Municipality 2015-06-01               57.02
## 6        Durham 03-32-010 Municipality 2015-10-01               38.72
## 7        Durham 03-32-010 Municipality 2015-03-01               43.10
## 8        Durham 03-32-010 Municipality 2015-07-01               41.65
## 9        Durham 03-32-010 Municipality 2015-11-01               43.55
## 10       Durham 03-32-010 Municipality 2015-04-01               49.68
## 11       Durham 03-32-010 Municipality 2015-08-01               44.70
## 12       Durham 03-32-010 Municipality 2015-12-01               48.75
```

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.

```r
#8
ashe_2015 <-scrape.it('01-11-010', '2015')

ashe_durham_2015 <- rbind(ashe_2015, durham_2015)

ggplot(ashe_durham_2015, aes(x = Month,
                             y = Max_Withdrawals_mgd,
                             color = Water.System)) +
  geom_point() +
  labs(title = "Water Usage (2015 Asheville)",
       y = "Max Withdrawals (MGD)")
```

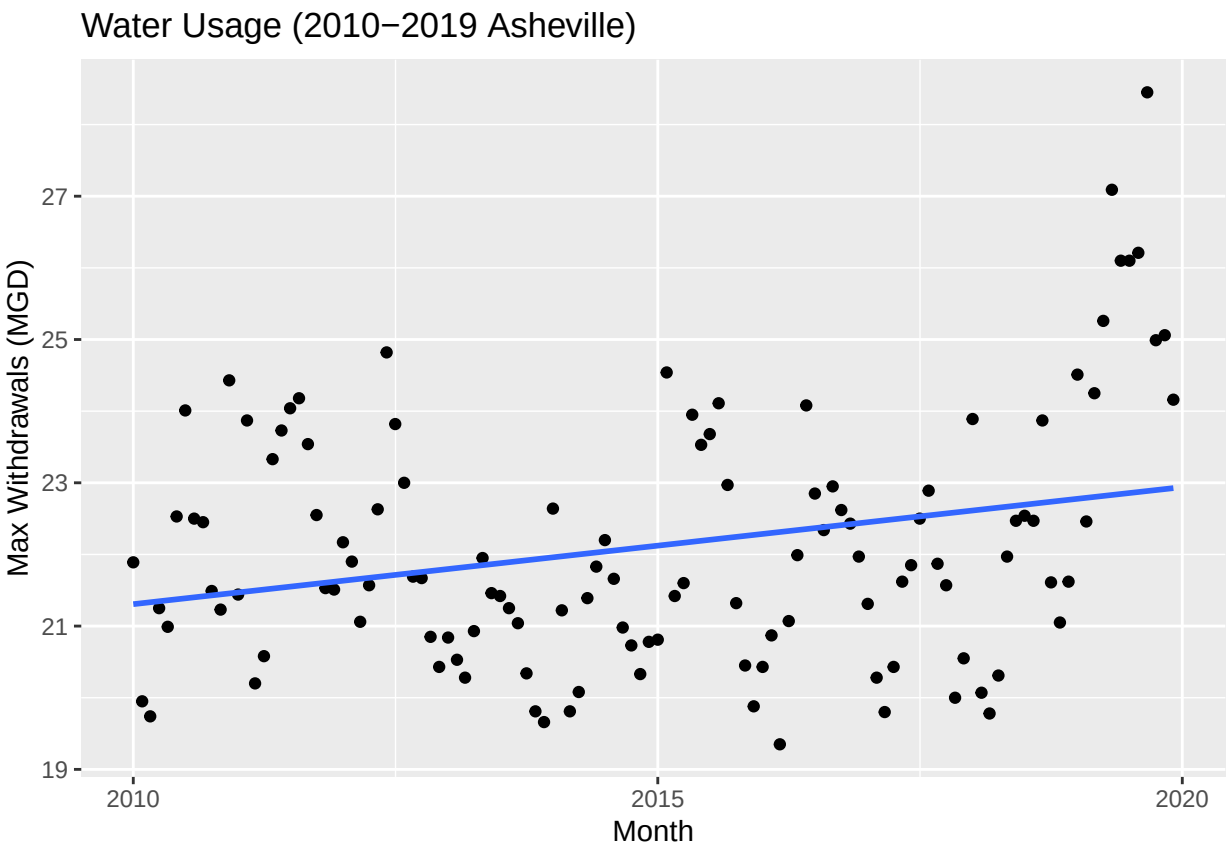


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2019.Add a smoothed line to the plot.

```r
#9

scrape.it <- function(pwsid, year) {

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

  #html_elemnts
  water.system.name <- 'div+ table tr:nth-child(1) td:nth-child(2)'
  pswid <- "td tr:nth-child(1) td:nth-child(5)"
```

```
  ownership <- "div+ table tr:nth-child(2) td:nth-child(4)"
  max.withdrawals.mgd <- "th~ td+ td"

  water_system <- website %>% html_nodes(water.system.name) %>% html_text()
  ID <- website %>% html_nodes(pswid) %>%  html_text()
  Ownership <- website %>% html_nodes(ownership) %>% html_text()
  max_withdrawals <- website %>% html_nodes(max.withdrawals.mgd) %>% html_text()

  #dataframe
  months <- c("Jan", "May", "Sep", "Feb", "Jun", "Oct",
              "Mar", "Jul", "Nov", "Apr", "Aug", "Dec")

  df <- data.frame("Water System" = rep(water_system, 12),
                   "PSWID" = rep(ID,12),
                   "Ownership" = rep(Ownership,12),
                   "Month" = my(paste(months,"-", year, sep ='')),
                   "Max_Withdrawals_mgd" = as.numeric(max_withdrawals))
  return(df)
}

years <- c(2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019)
df_fin <- data.frame()

for (i in years) {
  df <- scrape.it('01-11-010', i)
  ifelse(i == 2010,
         df_fin <- df,
         df_fin <- rbind(df_fin, df)
  )
  }

ggplot(df_fin, aes(x = Month, y = Max_Withdrawals_mgd)) +
  geom_point() +
  geom_smooth(method = lm, se = F) +
  labs(title = "Water Usage (2010-2019 Asheville)",
       y = "Max Withdrawals (MGD)")
```

```
## `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?

Asheville's water usage (based on Monthly Max Withdrawals) has increased over time.