Assignment 09: Data Scraping

Lehe, Xu

# 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] "D:/Documents/Environmental\_Data\_Analytics\_2022"

library(tidyverse)  
library(lubridate)  
  
#install.packages("rvest")  
library(rvest)

## Warning: 程辑包'rvest'是用R版本4.1.3 来建造的

# Set theme  
mytheme <- theme\_classic() +  
 theme(axis.text = element\_text(color = "black"),   
 legend.position = "top")  
theme\_set(mytheme)

1. 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  
webpage <- read\_html('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2020')  
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= ...

1. 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 <- 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()  
ownership <- webpage %>%   
 html\_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%   
 html\_text()  
max.withdrawals.mgd <- webpage %>%   
 html\_nodes("th~ td+ td") %>%   
 html\_text()

1. 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.)

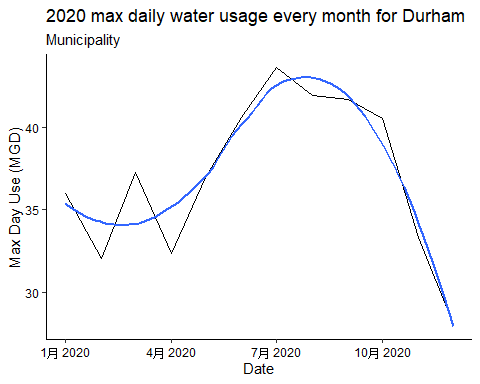
TIP: Use rep() to repeat a value when creating a dataframe.

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…

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

#4  
the\_df <- data.frame(  
 "water\_system\_name" = rep(water.system.name),  
 "pswid"= rep(pswid),  
 "ownership"=rep(ownership),  
 "max.withdrawals.mgd"= as.numeric(max.withdrawals.mgd),  
 "month"= c("1","5","9","2","6","10","3","7","11","4","8","12"),  
 "year"=rep(2020))  
  
the\_df <- the\_df %>%   
 mutate(Date = my(paste(month,"-",year)))  
#5  
ggplot(the\_df,aes(x=Date,y=max.withdrawals.mgd)) +   
 geom\_line() +   
 geom\_smooth(method="loess",se=FALSE) +  
 labs(title ="2020 max daily water usage every month for Durham",  
 subtitle = ownership,  
 y="Max Day Use (MGD)",  
 x="Date")

## `geom\_smooth()` using formula 'y ~ x'



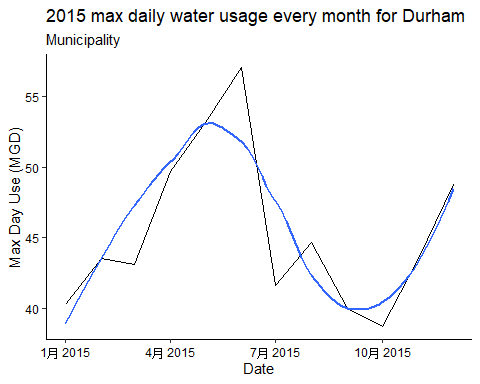
1. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data. **Be sure to modify the code to reflect the year and site scraped**.

#6.  
scrape.it <- function(pwsid,the\_year){  
   
webpage <- read\_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php?',   
 'pwsid=', pwsid, '&year=', the\_year))  
   
water.system.name <- 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()  
ownership <- webpage %>%   
 html\_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%   
 html\_text()  
max.withdrawals.mgd <- webpage %>%   
 html\_nodes("th~ td+ td") %>%   
 html\_text()  
  
the\_df <- data.frame(  
 "water\_system\_name" = rep(water.system.name),  
 "pswid"= rep(pswid),  
 "ownership"=rep(ownership),  
 "max.withdrawals.mgd"= as.numeric(max.withdrawals.mgd),  
 "month"= c("1","5","9","2","6","10","3","7","11","4","8","12"),  
 "year"=rep(the\_year)) %>%  
 mutate(Date = my(paste(month,"-",year)))  
  
return(the\_df)  
}

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

#7  
Durham\_df\_2015 <- scrape.it('03-32-010', 2015)  
view(Durham\_df\_2015)  
  
ggplot(Durham\_df\_2015,aes(x=Date,y=max.withdrawals.mgd)) +   
 geom\_line() +   
 geom\_smooth(method="loess",se=FALSE) +  
 labs(title = "2015 max daily water usage every month for Durham",  
 subtitle = ownership,  
 y="Max Day Use (MGD)",  
 x="Date")

## `geom\_smooth()` using formula 'y ~ x'



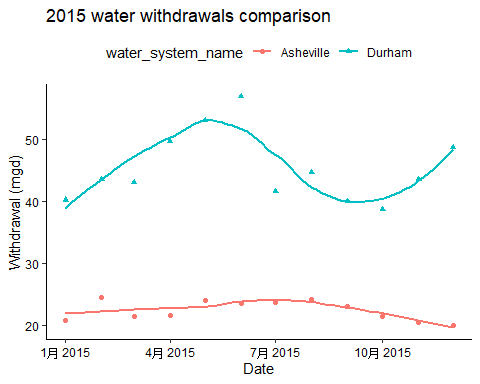
1. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares the Asheville to Durham’s water withdrawals.

#8  
Asheville\_df\_2015 <- scrape.it('01-11-010', 2015)  
view(Asheville\_df\_2015)  
  
join\_2015 <- full\_join(Asheville\_df\_2015,Durham\_df\_2015)

## Joining, by = c("water\_system\_name", "pswid", "ownership",  
## "max.withdrawals.mgd", "month", "year", "Date")

compare <-  
 ggplot(join\_2015, aes(x = Date, y = max.withdrawals.mgd, color = water\_system\_name, shape = water\_system\_name)) +  
 geom\_point() +  
 geom\_smooth(se=FALSE)+  
 labs(title = "2015 water withdrawals comparison",  
 y="Withdrawal (mgd)",  
 x="Date")  
print(compare)

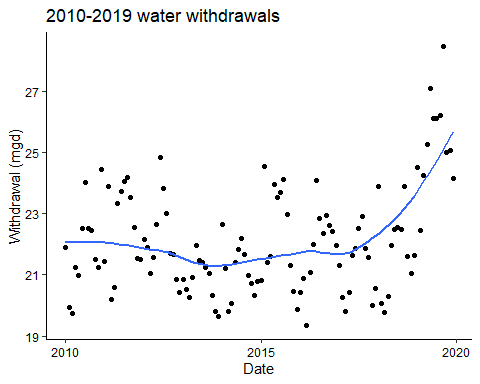
## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



1. 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 = c(2010:2019)  
  
All\_Asheville <- cross2('01-11-010', the\_years) %>%   
 map(lift(scrape.it)) %>%   
 bind\_rows()  
  
  
Asheville <-  
 ggplot(All\_Asheville, aes(x = Date, y = max.withdrawals.mgd)) +  
 geom\_point() +  
 geom\_smooth(se=FALSE)+  
 labs(title = "2010-2019 water withdrawals",  
 y="Withdrawal (mgd)",  
 x="Date")  
print(Asheville)

## `geom\_smooth()` using method = 'loess' and 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?

Yes, Asheville have a trend in water usage over time. From 2010-2015, the water usage was gradually decreasing, but after 2015 the water usage started increasing. Especially, afte 2018, water usage increased rapidly. Overall, water usage shows an increasing trend over time.