Assignment 09: Data Scraping

Xuening Tang

## OVERVIEW

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

## Directions

1. Rename this file <FirstLast>\_A09\_DataScraping.Rmd (replacing <FirstLast> with your first and last name).
2. Change “Student Name” on line 3 (above) with your name.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.

## 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] "/Users/xueningtang/Desktop/R lab/EDA-Fall2022/Assignments"

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.1.2

## Warning: package 'tibble' was built under R version 4.1.2

## Warning: package 'tidyr' was built under R version 4.1.2

## Warning: package 'readr' was built under R version 4.1.2

## Warning: package 'dplyr' was built under R version 4.1.2

library(rvest)

## Warning: package 'rvest' was built under R version 4.1.2

library(lubridate)  
  
my.theme <- theme\_classic(base\_size = 12) +  
 theme(axis.text = element\_text(color = "black"),   
 legend.position = "top")  
theme\_set(my.theme)

1. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham’s 2021 Municipal Local Water Supply Plan (LWSP):

* Navigate to <https://www.ncwater.org/WUDC/app/LWSP/search.php>
* 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=2021>

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

#2  
webpage.1 <- read\_html('https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2021')  
webpage.1

## {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:
* Maximum Daily Use (MGD) - for each month

In the code chunk below scrape these values, assigning them to four 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 (represented as strings), with the first value being “27.6400”.

#3  
water.system.name <- webpage.1 %>%   
 html\_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%   
 html\_text()  
   
pswid <- webpage.1 %>%   
 html\_nodes("td tr:nth-child(1) td:nth-child(5)") %>%   
 html\_text()  
   
ownership <- webpage.1 %>%   
 html\_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%   
 html\_text()  
   
max.withdrawals.mgd <- webpage.1 %>%   
 html\_nodes("th~ td+ td") %>%   
 html\_text()

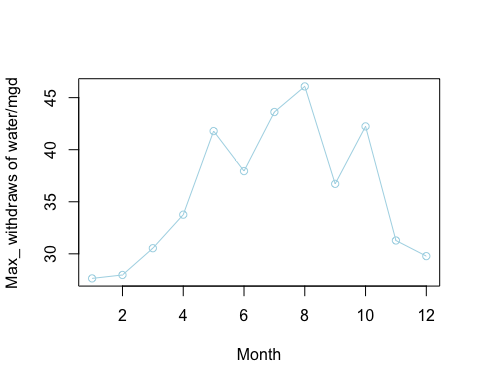
1. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

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 chronological order. You can overcome this by creating a month column manually assigning values in the order the data are scraped: “Jan”, “May”, “Sept”, “Feb”, etc…

1. Create a line plot of the maximum daily withdrawals across the months for 2021

#4  
Max\_withdraws\_mgd <- data.frame(max.withdrawals.mgd)  
Max\_final <- Max\_withdraws\_mgd[c(1,4,7,10,2,5,8,11,3,6,9,12),]  
  
df\_water <- data.frame("max.withdrawals.mgd" = Max\_final,  
 "Month" = rep(1:12),  
 "Year" = rep(2021,12))%>%  
 mutate(Water\_System\_Name = !!water.system.name,  
 PSWID = !!pswid,  
 Ownership = !!ownership,  
 Date = my(paste(Month,"-",Year)))  
  
  
  
#5  
  
plot(df\_water$max.withdrawals.mgd,type = "o", col = "light blue", xlab = "Month", ylab = "Max\_ withdraws of water/mgd")



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 (pwsid) scraped**.

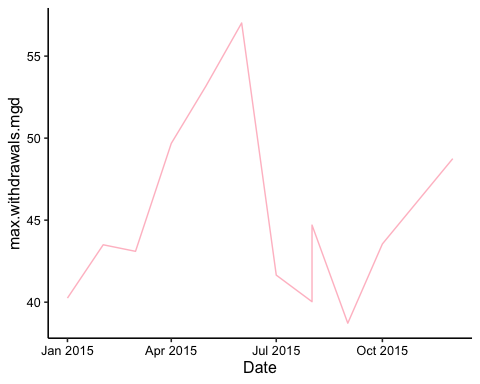
#6.  
base\_url <- 'https://www.ncwater.org/WUDC/app/LWSP/report.php'  
my\_PWSID <- 'pwsid=03-43-015'  
my\_Year <- 'year=2017'  
my\_scrape\_url <- paste0(base\_url, '?', my\_PWSID, '&', my\_Year)  
print(my\_scrape\_url)

## [1] "https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-43-015&year=2017"

my\_website <- read\_html(my\_scrape\_url)  
my\_system\_name\_tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'  
my\_pwsid\_tag <- 'td tr:nth-child(1) td:nth-child(5)'  
my\_ownership\_tag <- 'div+ table tr:nth-child(2) td:nth-child(4)'  
my\_max\_withdraw\_tag <- 'th~ td+ td'  
  
my\_system\_name <- my\_website %>% html\_nodes(my\_system\_name\_tag) %>% html\_text()  
my\_pwsid\_name <- my\_website %>% html\_nodes(my\_pwsid\_tag) %>%html\_text()  
my\_ownership <- my\_website %>% html\_nodes(my\_ownership\_tag) %>% html\_text()  
my\_max\_withdrawals <- my\_website %>% html\_nodes(my\_max\_withdraw\_tag)%>%html\_text()  
  
  
my\_scrape <- function(my\_PWSID,my\_Year){  
 my\_website <- read\_html(paste0('https://www.ncwater.org/WUDC/app/LWSP/report.php'  
 , '?pwsid=', my\_PWSID, '&year=', my\_Year))  
  
 my\_system\_name\_tag <- 'div+ table tr:nth-child(1) td:nth-child(2)'  
 my\_pwsid\_tag <- 'td tr:nth-child(1) td:nth-child(5)'  
 my\_ownership\_tag <- 'div+ table tr:nth-child(2) td:nth-child(4)'  
 my\_max\_withdraw\_tag <- 'th~ td+ td'  
   
 my\_system\_name <- my\_website %>% html\_nodes(my\_system\_name\_tag) %>% html\_text()  
 my\_pwsid\_name <- my\_website %>% html\_nodes(my\_pwsid\_tag) %>%html\_text()  
 my\_ownership <- my\_website %>% html\_nodes(my\_ownership\_tag) %>% html\_text()  
 my\_max\_withdrawals <- my\_website %>% html\_nodes(my\_max\_withdraw\_tag)%>%html\_text()  
   
 df\_water <- data.frame("max.withdrawals.mgd" = as.numeric(my\_max\_withdrawals),  
 "Year" = rep(my\_Year,12))%>%  
 mutate(Water\_System\_Name = !!my\_system\_name,  
 PWSID = !!my\_pwsid\_name,  
 Ownership = !!my\_ownership)  
   
return(df\_water)  
}

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

#7  
durh\_df <- my\_scrape('03-32-010','2015')  
view(durh\_df)  
  
df\_durh <- durh\_df %>%   
 mutate(Month = c(1,5,8,2,6,9,3,7,10,4,8,12),  
 Date = my(paste(Month,"-",Year)))  
durh\_final <- df\_durh[c(1,4,7,10,2,5,8,11,3,6,9,12),]  
  
ggplot(df\_durh,aes(x=Date,y=max.withdrawals.mgd)) +   
 geom\_line(color="pink")

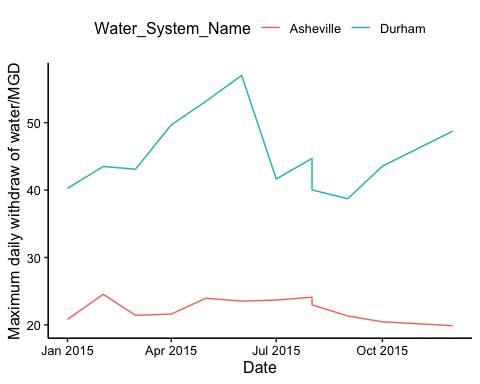


1. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares Asheville’s to Durham’s water withdrawals.

#8  
df\_ashe <- my\_scrape('01-11-010','2015')  
view(df\_ashe)  
ashe\_df <- df\_ashe %>%   
 mutate(Month = c(1,5,8,2,6,9,3,7,10,4,8,12),  
 Date = my(paste(Month,"-",Year)))  
ashe\_final <- ashe\_df[c(1,4,7,10,2,5,8,11,3,6,9,12),]  
  
df\_combine <- full\_join(ashe\_final,durh\_final)

## Joining, by = c("max.withdrawals.mgd", "Year", "Water\_System\_Name", "PWSID",  
## "Ownership", "Month", "Date")

water.plot<-  
 ggplot(df\_combine,aes(x=Date,y=max.withdrawals.mgd, color=Water\_System\_Name)) +   
 geom\_line()+  
 ylab("Maximum daily withdraw of water/MGD")+  
 my.theme  
print(water.plot)



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.

TIP: See Section 3.2 in the “09\_Data\_Scraping.Rmd” where we apply “map2()” to iteratively run a function over two inputs. Pipe the output of the map2() function to bindrows() to combine the dataframes into a single one.

#9  
my\_years = rep(2010:2019)  
the\_pwsid = '01-11-010'

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