Assignment 10: Data Scraping

Keanu Valibia

Contents

List of Figures

OVERVIEW

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

Directions

1. Rename this file `<FirstLast>_A10_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 your code is tidy; use line breaks to ensure your code fits in the knitted output.
5. Be sure to **answer the questions** in this assignment document.
6. When you have completed the assignment, **Knit** the text and code into a single PDF file.

Set up

1. Set up your session:

- Load the packages `tidyverse`, `rvest`, and any others you end up using.
- Check your working directory

```
knitr::opts_chunk$set(tidy.opts=list(width.cutoff=45), tidy=TRUE)
```

```r
# 1
library(tidyverse)
library(lubridate)
# install.packages('rvest')
library(rvest)
library(here)

here()
```

```
## [1] "/home/guest/R/R Projects/EDA_Spring2024"
```

```r
myTheme <- theme_classic(base_size = 11) + theme(axis.text = element_text(color = "black"),
    axis.line = element_line(color = "black"),
    panel.background = element_rect(fill = "#EDE6E3"),
    panel.grid.major = element_line(color = "#36382E",
        linetype = "dotted"), plot.title = element_text(size = 15),
    axis.title.x = element_text(size = 13), axis.title.y = element_text(size = 13),
    legend.position = "right")

theme_set(myTheme)
```

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham's 2022 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=2022

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

```r
# 2

webpage <- read_html("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2022")
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 "1. System Information" section:
- Water system name
- PWSID
- Ownership

- From the "3. Water Supply Sources" section:
- Maximum Day 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)".

```
# 3

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

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

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

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

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

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

```
monthly_max_day_usage <- webpage %>%
    html_nodes("th~ td+ td") %>%
    html_text()
monthly_max_day_usage
```

```
##  [1] "36.1000" "43.4200" "52.4900" "30.5000" "42.5900" "34.8800" "39.9100"
##  [8] "43.3200" "32.5300" "34.6600" "41.8000" "37.5300"
```

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 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... Or, you could scrape month values from the web page...

5. Create a line plot of the maximum daily withdrawals across the months for 2022

```r
# 4

durham2022.df <- data.frame(Month = as.factor(c("01",
    "05", "09", "02", "06", "10", "03", "07",
    "11", "04", "08", "12")), Year = rep(2022,
    12), `Water System` = water_system, PWSID = PWSID,
    Ownership = ownership, `Monthly Max Usage` = as.numeric(monthly_max_day_usage))

durham2022.df <- arrange(durham2022.df, by_group = Month)

# 5

max.daily.withdrawals.plot <- ggplot(durham2022.df,
    aes(x = Month, y = Monthly.Max.Usage, group = 1)) +
    geom_line() + geom_point() + labs(title = "Monthly Max Usage (MGD) Over Time",
    subtitle = "Year: 2022", caption = "Source: www.ncwater.org",
    ) + ylab("Monthly Max Usage (MGD)") + xlab("Month")

max.daily.withdrawals.plot
```

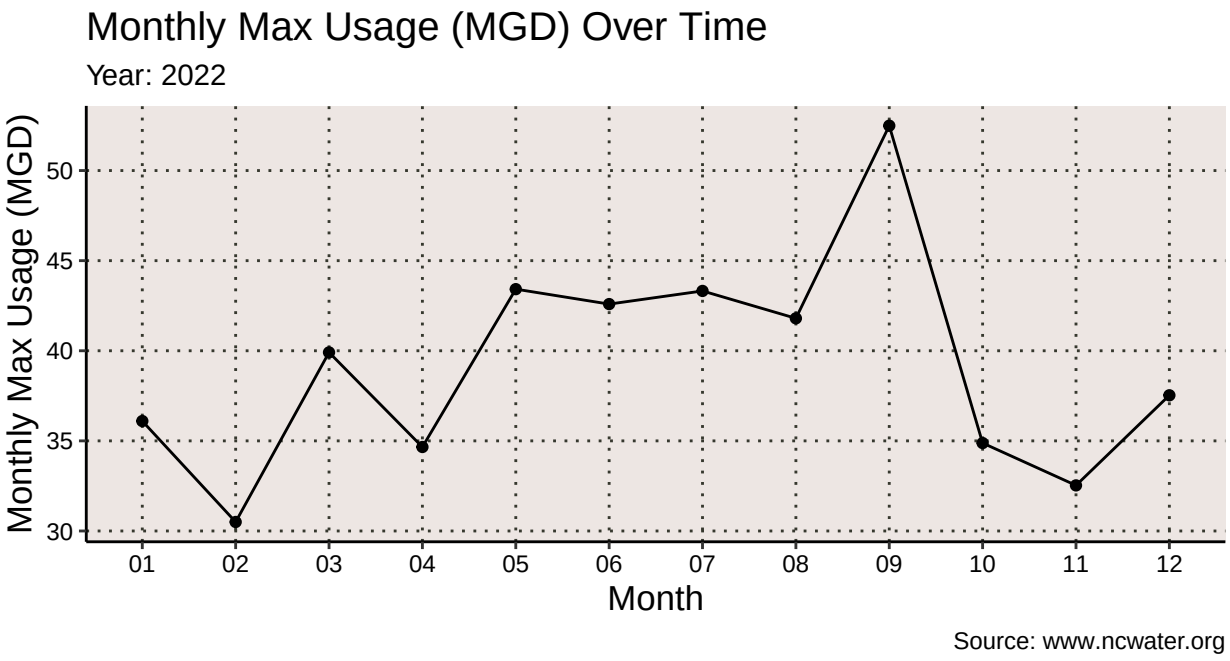

Figure 1: Durham 2022 - Monthly Max Withdrawls

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 (pwsid) scraped**.

```r
# 6.

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

```
    # Get site content
    site.url <- read_html(paste0("https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=",
        pwsid.tag, "&year=", year))

    # Set variables
    water.system.tag <- "div+ table tr:nth-child(1) td:nth-child(2)"
    pwsid.tag <- "td tr:nth-child(1) td:nth-child(5)"
    ownership.tag <- "div+ table tr:nth-child(2) td:nth-child(4)"
    usage.tag <- "th~ td+ td"

    # Scrape data
    water.system.scrape <- site.url %>%
        html_nodes(water.system.tag) %>%
        html_text()
    pwsid.scrape <- site.url %>%
        html_nodes(pwsid.tag) %>%
        html_text()
    ownership.scrape <- site.url %>%
        html_nodes(ownership.tag) %>%
        html_text()
    usage.scrape <- site.url %>%
        html_nodes(usage.tag) %>%
        html_text()

    water.system.scrape <- water.system.scrape[1]
    pwsid.scrape <- pwsid.scrape[1]
    ownership.scrape <- ownership.scrape[1]

    # Convert to dataframe
    df.withdrawals <- data.frame(Month = as.factor(c("01",
        "05", "09", "02", "06", "10", "03", "07",
        "11", "04", "08", "12")), Year = rep(year,
        12), Max.Daily.Withdrawals = as.numeric(usage.scrape)) %>%
        mutate(WaterSystem = water.system.scrape,
            PWSID = pwsid.scrape, Ownership = ownership.scrape,
            Date = my(paste(Month, "-", Year)))

    # Pause
    Sys.sleep(1)

    # Return dataframe
    return(df.withdrawals)
}
```

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

```
# 7

durham2015.df <- scrape.function(2015, "03-32-010")
view(durham2015.df)
```

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 Asheville's to Durham's water withdrawals.

```
# 8

asheville2015.df <- scrape.function(2015, "01-11-010")

combined.df <- rbind(durham2015.df, asheville2015.df)

combined.plot <- ggplot(combined.df, aes(x = Month,
    y = Max.Daily.Withdrawals, group = WaterSystem,
    color = WaterSystem)) + geom_line() + geom_point() +
    labs(title = "Max Usage (MGD) in Durham and Asheville",
        subtitle = "Year: 2015", caption = "Source: www.ncwater.org",
        color = "Water System") + ylab("Max Dailiy Usage (MGD)") +
    xlab("Month")

combined.plot
```

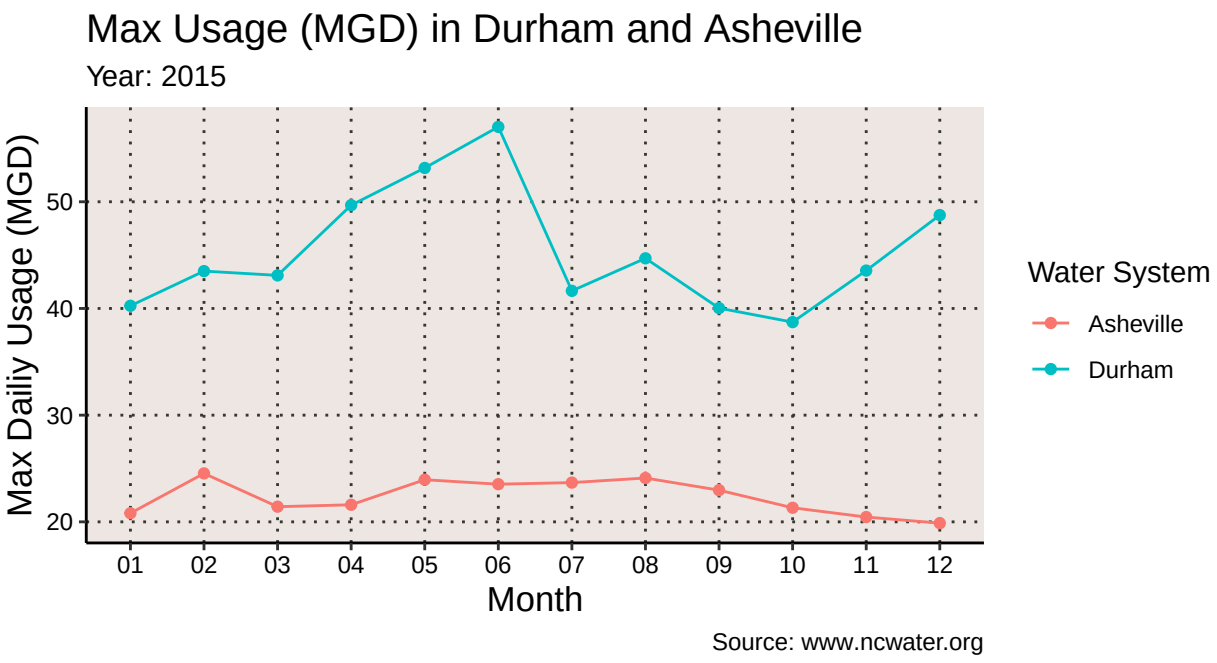

Figure 2: Durham & Asheville 2015 - Monthly Max Withdrawls

9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2010 thru 2021.Add a smoothed line to the plot (method = 'loess').

   TIP: See Section 3.2 in the "10_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

years.subset <- rep(2010:2015)

asheville.2010.2015.df <- map2(years.subset, "01-11-010",
    scrape.function) %>%
    bind_rows()

asheville.2010.2015.plot <- ggplot(asheville.2010.2015.df,
    aes(x = Month, y = Max.Daily.Withdrawals,
        group = as.factor(Year), color = as.factor(Year))) +
    geom_line() + labs(title = "Max Usage (MGD) in Asheville",
    subtitle = "Years: 2010 - 2015", caption = "Source: www.ncwater.org",
    color = "Year") + scale_color_manual(values = c("#5A8D3B",
    "#197278", "#FFAB00", "#C44536", "#772E25",
    "#490E25")) + ylab("Max Dailiy Usage (MGD)") +
    xlab("Month")

asheville.2010.2015.plot
```

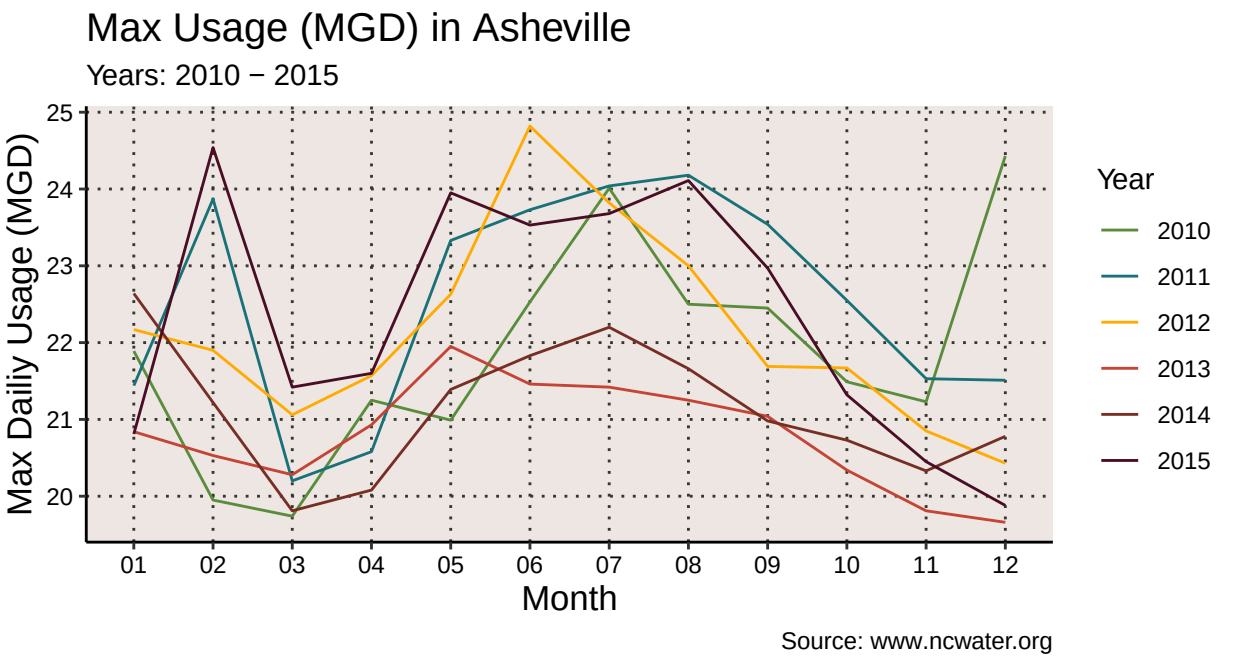

Figure 3: Asheville 2010 through 2015 - Monthly Max Withdrawls

Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time? > Answer: Over the years, water usage has fluctuated up and down. However, the usage within a year tends to decrease in the months of May and April, increase over the sumer, and then fall once again towards the Fall and Winter.