Assignment 10: Data Scraping

Thomas Hancock

Total points:

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on time series analysis.

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., "Salk_A06_GLMs_Week1.Rmd") prior to submission.

The completed exercise is due on Tuesday, April 7 at 1:00 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

```
getwd()
```

```
## [1] "C:/Users/thoma/Thomas/2018 Grad School/Duke MEM/ENV 872/Environmental_Data_Analytics_2020/Assig
```

```
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3

## Warning: package 'tidyr' was built under R version 3.6.3

## Warning: package 'dplyr' was built under R version 3.6.3

## Warning: package 'forcats' was built under R version 3.6.3
```

```
library(rvest)
```

```
## Warning: package 'xml2' was built under R version 3.6.3
```

```
library(viridis)
library(ggrepel)
```

```
## Warning: package 'ggrepel' was built under R version 3.6.3
```

```
# Set theme
mytheme <- theme_classic() +
```

```r
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
theme_set(mytheme)
```

2. Indicate the EPA impaired waters website (https://www.epa.gov/nutrient-policy-data/waters-assessed-impaired-due-nutrient-related-causes) as the URL to be scraped.

```r
# Specify website to be scraped
url <- "https://www.epa.gov/nutrient-policy-data/waters-assessed-impaired-due-nutrient-related-causes"

# Reading the HTML code from the website
webpage <- read_html(url)
```

3. Scrape the Rivers table, with every column except year. Then, turn it into a data frame.

```r
State <- webpage %>% html_nodes("table:nth-child(8) td:nth-child(1)") %>% html_text()
Rivers.Assessed.mi2 <- webpage %>% html_nodes("table:nth-child(8) td:nth-child(2)") %>%
  html_text()
Rivers.Assessed.percent <- webpage %>% html_nodes("table:nth-child(8) td:nth-child(3)") %>%
  html_text()
Rivers.Impaired.mi2 <- webpage %>% html_nodes("table:nth-child(8) td:nth-child(4)") %>%
  html_text()
Rivers.Impaired.percent <- webpage %>% html_nodes("table:nth-child(8) td:nth-child(5)") %>%
  html_text()
Rivers.Impaired.percent.TMDL <- webpage %>%
  html_nodes("table:nth-child(8) td:nth-child(6)") %>% html_text()

Rivers <- data.frame(State, Rivers.Assessed.mi2, Rivers.Assessed.percent, Rivers.Impaired.mi2,
                     Rivers.Impaired.percent, Rivers.Impaired.percent.TMDL)
```

4. Use `str_replace` to remove non-numeric characters from the numeric columns.
5. Set the numeric columns to a numeric class and verify this using `str`.

```r
# 4
# Use str_replace to remove non-numeric characters
Rivers$Rivers.Assessed.mi2 <- str_replace(Rivers$Rivers.Assessed.mi2, pattern = "([,])",
                                          replacement = "")
Rivers$Rivers.Assessed.percent <- str_replace(Rivers$Rivers.Assessed.percent,
                                              pattern = "([%])", replacement = "")
Rivers$Rivers.Assessed.percent <- str_replace(Rivers$Rivers.Assessed.percent,
                                              pattern = "([*])", replacement = "")
Rivers$Rivers.Impaired.mi2 <- str_replace(Rivers$Rivers.Impaired.mi2, pattern = "([,])",
                                          replacement = "")
Rivers$Rivers.Impaired.percent <- str_replace(Rivers$Rivers.Impaired.percent,
                                              pattern = "([%])", replacement = "")
Rivers$Rivers.Impaired.percent.TMDL <- str_replace(Rivers$Rivers.Impaired.percent.TMDL,
                                              pattern = "([%])", replacement = "")
Rivers$Rivers.Impaired.percent.TMDL <- str_replace(Rivers$Rivers.Impaired.percent.TMDL,
                                              pattern = "([±])", replacement = "")

# 5
str(Rivers)
```

```
## 'data.frame':    50 obs. of  6 variables:
##  $ State                       : Factor w/ 50 levels "Alabama","Alaska",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Rivers.Assessed.mi2         : chr  "10538" "602" "2764" "9979" ...
```

```
##  $ Rivers.Assessed.percent     : chr  "14" "0" "3" "11" ...
##  $ Rivers.Impaired.mi2         : chr  "1146" "15" "144" "1440" ...
##  $ Rivers.Impaired.percent     : chr  "11" "2" "5" "14" ...
##  $ Rivers.Impaired.percent.TMDL: chr  "53" "100" "6" "2" ...
```

```
Rivers$Rivers.Assessed.mi2 <- as.numeric(Rivers$Rivers.Assessed.mi2)
Rivers$Rivers.Assessed.percent <- as.numeric(Rivers$Rivers.Assessed.percent)
Rivers$Rivers.Impaired.mi2 <- as.numeric(Rivers$Rivers.Impaired.mi2)
Rivers$Rivers.Impaired.percent <- as.numeric(Rivers$Rivers.Impaired.percent)
Rivers$Rivers.Impaired.percent.TMDL <- as.numeric(Rivers$Rivers.Impaired.percent.TMDL)

str(Rivers)
```

```
## 'data.frame':    50 obs. of  6 variables:
##  $ State                       : Factor w/ 50 levels "Alabama","Alaska",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Rivers.Assessed.mi2         : num  10538 602 2764 9979 32803 ...
##  $ Rivers.Assessed.percent     : num  14 0 3 11 16 56 41 100 20 19 ...
##  $ Rivers.Impaired.mi2         : num  1146 15 144 1440 13350 ...
##  $ Rivers.Impaired.percent     : num  11 2 5 14 41 0 0 88 53 9 ...
##  $ Rivers.Impaired.percent.TMDL: num  53 100 6 2 NA 14 73 37 NA 78 ...
```

6. Scrape the Lakes table, with every column except year. Then, turn it into a data frame.

```
State <- webpage %>% html_nodes("table:nth-child(14) td:nth-child(1)") %>% html_text()
Lakes.Assessed.mi2 <- webpage %>% html_nodes("table:nth-child(14) td:nth-child(2)") %>%
  html_text()
Lakes.Assessed.percent <- webpage %>% html_nodes("table:nth-child(14) td:nth-child(3)") %>%
  html_text()
Lakes.Impaired.mi2 <- webpage %>% html_nodes("table:nth-child(14) td:nth-child(4)") %>%
  html_text()
Lakes.Impaired.percent <- webpage %>% html_nodes("table:nth-child(14) td:nth-child(5)") %>%
  html_text()
Lakes.Impaired.percent.TMDL <- webpage %>%
  html_nodes("table:nth-child(14) td:nth-child(6)") %>% html_text()

Lakes <- data.frame(State, Lakes.Assessed.mi2, Lakes.Assessed.percent,
                    Lakes.Impaired.mi2, Lakes.Impaired.percent, Lakes.Impaired.percent.TMDL)
```

7. Filter out the states with no data.
8. Use `str_replace` to remove non-numeric characters from the numeric columns.
9. Set the numeric columns to a numeric class and verify this using `str`.

```
# 7
# Filter out states with no data
Lakes <- Lakes %>%
  filter(State != "Hawaii" & State != "Pennsylvania")



# 8
# Use str_replace to remove non-numeric characters
Lakes$Lakes.Assessed.mi2 <- str_replace(Lakes$Lakes.Assessed.mi2, pattern = "([,])",
                                        replacement = "")
Lakes$Lakes.Assessed.mi2 <- str_replace(Lakes$Lakes.Assessed.mi2, pattern = "([,])",
                                        replacement = "") # This is included twice because some state
Lakes$Lakes.Assessed.percent <- str_replace(Lakes$Lakes.Assessed.percent,
```

```
                                                 pattern = "([%])", replacement = "")
Lakes$Lakes.Assessed.percent <- str_replace(Lakes$Lakes.Assessed.percent,
                                                 pattern = "([*])", replacement = "")
Lakes$Lakes.Impaired.mi2 <- str_replace(Lakes$Lakes.Impaired.mi2, pattern = "([,])",
                                             replacement = "")
Lakes$Lakes.Impaired.percent <- str_replace(Lakes$Lakes.Impaired.percent,
                                                 pattern = "([%])", replacement = "")
Lakes$Lakes.Impaired.percent.TMDL <- str_replace(Lakes$Lakes.Impaired.percent.TMDL,
                                                 pattern = "([%])", replacement = "")
Lakes$Lakes.Impaired.percent.TMDL <- str_replace(Lakes$Lakes.Impaired.percent.TMDL,
                                                 pattern = "([±])", replacement = "")


# 9
str(Lakes)
```

```
## 'data.frame':    48 obs. of  6 variables:
##  $ State                      : Factor w/ 50 levels "Alabama","Alaska",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Lakes.Assessed.mi2         : chr  "430.976" "5981" "114976" "64778" ...
##  $ Lakes.Assessed.percent     : chr  "88" "0" "34" "13" ...
##  $ Lakes.Impaired.mi2         : chr  "81740" "1137" "4895" "6513" ...
##  $ Lakes.Impaired.percent     : chr  "19" "19" "4" "10" ...
##  $ Lakes.Impaired.percent.TMDL: chr  "53" "73" "9" "71" ...
```

```
Lakes$Lakes.Assessed.mi2 <- as.numeric(Lakes$Lakes.Assessed.mi2)
Lakes$Lakes.Assessed.percent <- as.numeric(Lakes$Lakes.Assessed.percent)
Lakes$Lakes.Impaired.mi2 <- as.numeric(Lakes$Lakes.Impaired.mi2)
Lakes$Lakes.Impaired.percent <- as.numeric(Lakes$Lakes.Impaired.percent)
Lakes$Lakes.Impaired.percent.TMDL <- as.numeric(Lakes$Lakes.Impaired.percent.TMDL)

str(Lakes)
```

```
## 'data.frame':    48 obs. of  6 variables:
##  $ State                      : Factor w/ 50 levels "Alabama","Alaska",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Lakes.Assessed.mi2         : num  431 5981 114976 64778 1051246 ...
##  $ Lakes.Assessed.percent     : num  88 0 34 13 50 95 47 100 54 82 ...
##  $ Lakes.Impaired.mi2         : num  81740 1137 4895 6513 473954 ...
##  $ Lakes.Impaired.percent     : num  19 19 4 10 45 7 12 88 82 2 ...
##  $ Lakes.Impaired.percent.TMDL: num  53 73 9 71 NA 0 7 69 NA 20 ...
```

10. Join the two data frames with a `full_join`.

```
# Do a full join on the two tables
RiversAndLakes <- full_join(Rivers, Lakes, by = "State")
```

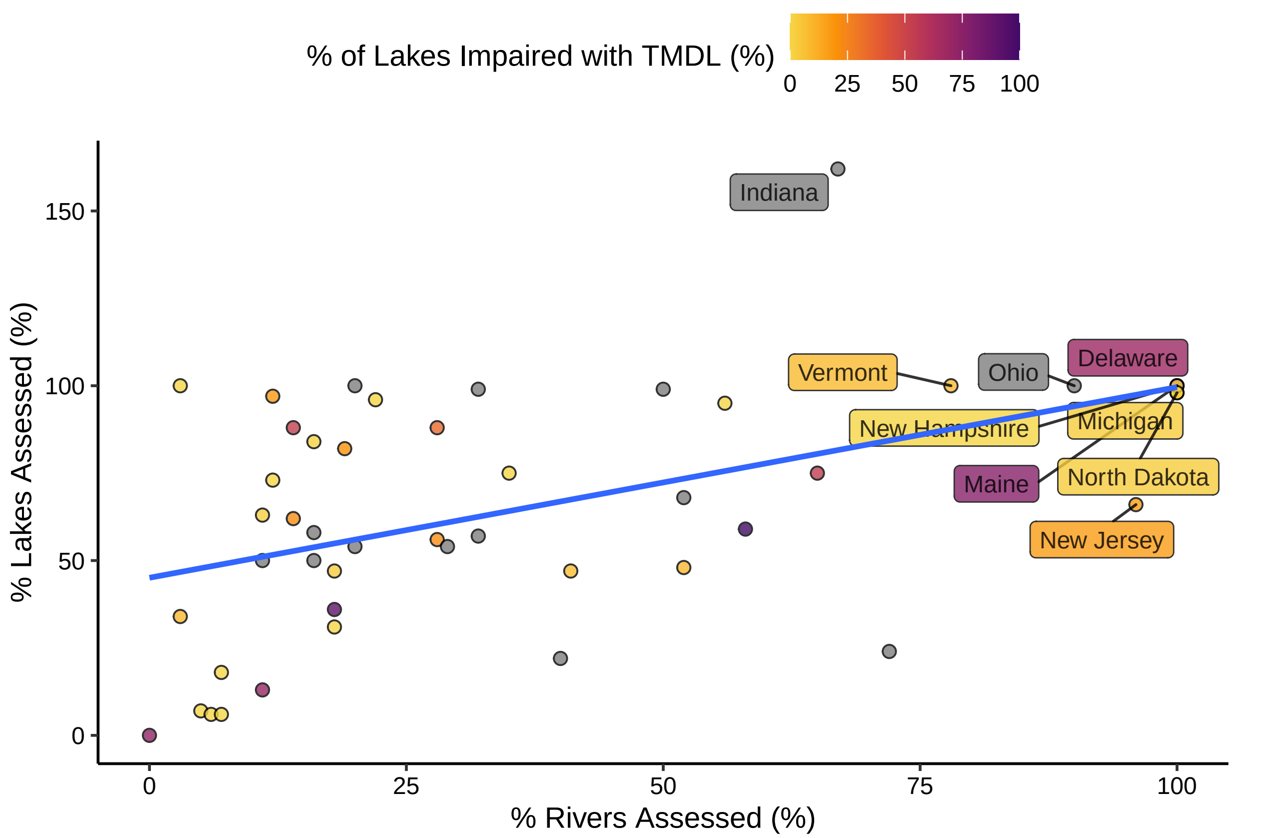
11. Create one graph that compares the data for lakes and/or rivers. This option is flexible; choose a relationship (or relationships) that seem interesting to you, and think about the implications of your findings. This graph should be edited so it follows best data visualization practices.

(You may choose to run a statistical test or add a line of best fit; this is optional but may aid in your interpretations)

```
# Make a point plot of % Lakes Assessed vs. % Rivers Assessed for each state
ggplot(RiversAndLakes, aes(x = Rivers.Assessed.percent, y = Lakes.Assessed.percent,
                           fill = Lakes.Impaired.percent.TMDL)) +

  geom_point(shape = 21, size = 2, alpha = 0.8, color = "black") +
```

```r
  scale_fill_viridis_c(option = "inferno", begin = 0.2, end = 0.9, direction = -1) +

  labs(x = "% Rivers Assessed (%)",
       y = "% Lakes Assessed (%)",
       fill = "% of Lakes Impaired with TMDL (%)") +

  # Lbel the states with >75% of rivers assessed (and the outlier, Indiana, that has >100%)
   geom_label_repel(data = subset(RiversAndLakes, Rivers.Assessed.percent > 75 |
                                   Lakes.Assessed.percent > 100),
                    aes(label = State), nudge_x = -5, nudge_y = -5, size = 3, alpha = 0.8) +

  geom_smooth(method = "lm", se = FALSE) # Add a regression line
```

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

## Warning: Removed 2 rows containing non-finite values (stat_smooth).

## Warning: Removed 2 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing missing values (geom_label_repel).
```



12. Summarize the findings that accompany your graph. You may choose to suggest further research or data collection to help explain the results.

    It seems like there is a rough trend among states that, as the % of rivers assessed increases, the % of lakes assessed increases as well (in other words, states that test more rivers are likely to test more lakes). That said, there are many states that test most of their lakes without testing as many of the rivers. In fact, there seems to be a preference among states for testing lakes over rivers (most states have tested more of their lakes than rivers). Of states with high percentages of lakes and rivers tested, there is a mix of states with the % of lakes impaired with TMDL. Finally, Indiana seems to have a data entry error, as they are shown as having assessed ~160% of their lakes.