Assignment 10: Data Scraping

Jack Eynon

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

```r
getwd()
```

```
## [1] "/Users/jackeynon/Courses/EnvDataAnalytics/Environmental_Data_Analytics_2020/Assignments"
```

```r
library(tidyverse)
library(rvest)
library(ggrepel)

mytheme <- theme_classic() +
  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
url <- "https://www.epa.gov/nutrient-policy-data/waters-assessed-impaired-due-nutrient-related-causes"
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.mi <- 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.mi <- 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_te

Rivers <- data.frame(State, Rivers.Assessed.mi, Rivers.Assessed.percent, Rivers.Impaired.mi,
                     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`.

```
# 4
Rivers$Rivers.Assessed.mi <- str_replace(Rivers$Rivers.Assessed.mi,
                                         pattern = "([,])", replacement = "")
Rivers$Rivers.Assessed.percent <- str_replace(Rivers$Rivers.Assessed.percent,
                                         pattern = "([%])", replacement = "")
Rivers$Rivers.Impaired.mi <- str_replace(Rivers$Rivers.Impaired.mi,
                                         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.mi          : chr  "10538" "602" "2764" "9979" ...
##  $ Rivers.Assessed.percent     : chr  "14" "0" "3" "11" ...
##  $ Rivers.Impaired.mi          : chr  "1146" "15" "144" "1440" ...
##  $ Rivers.Impaired.percent     : chr  "11" "2" "5" "14" ...
##  $ Rivers.Impaired.percent.TMDL: chr  "53" "100" "6" "2" ...
```

```
Rivers$Rivers.Assessed.mi <- as.numeric(Rivers$Rivers.Assessed.mi)
Rivers$Rivers.Assessed.percent <- as.numeric(Rivers$Rivers.Assessed.percent)
```

```
## Warning: NAs introduced by coercion
```

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

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.acres <- 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.acres <- 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_te

Lakes <- data.frame(State, Lakes.Assessed.acres, Lakes.Assessed.percent, Lakes.Impaired.acres, Lakes.Imp
```

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
Lakes %>% filter(Lakes.Impaired.acres == "No data")
```

```
##          State Lakes.Assessed.acres Lakes.Assessed.percent Lakes.Impaired.acres
## 1       Hawaii              No data                No data              No data
## 2 Pennsylvania              No data                No data              No data
##   Lakes.Impaired.percent Lakes.Impaired.percent.TMDL
## 1                No data                     No data
## 2                No data                     No data
```

```
Lakes <- Lakes %>% filter(State != "Hawaii" & State != "Pennsylvania")
# 8
Lakes$Lakes.Assessed.acres <- str_replace(Lakes$Lakes.Assessed.acres,
                                          pattern = "([,])", replacement = "")
Lakes$Lakes.Assessed.acres <- str_replace(Lakes$Lakes.Assessed.acres, ## removing misplaced decimal
                                          pattern = "([.])", replacement = "")
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$Lakes.Impaired.acres <- str_replace(Lakes$Lakes.Impaired.acres,
                                          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.acres       : chr  "430976" "5981" "114976" "64778" ...
##  $ Lakes.Assessed.percent     : chr  "88" "0" "34" "13" ...
##  $ Lakes.Impaired.acres       : chr  "81740" "1137" "4895" "6513" ...
##  $ Lakes.Impaired.percent     : chr  "19" "19" "4" "10" ...
##  $ Lakes.Impaired.percent.TMDL: chr  "53" "73" "9" "71" ...
```

```
Lakes$Lakes.Assessed.acres <- as.numeric(Lakes$Lakes.Assessed.acres)
```

```
## Warning: NAs introduced by coercion
```

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

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

```r
Rivers.and.Lakes <- full_join(Rivers, Lakes, by = "State")
```

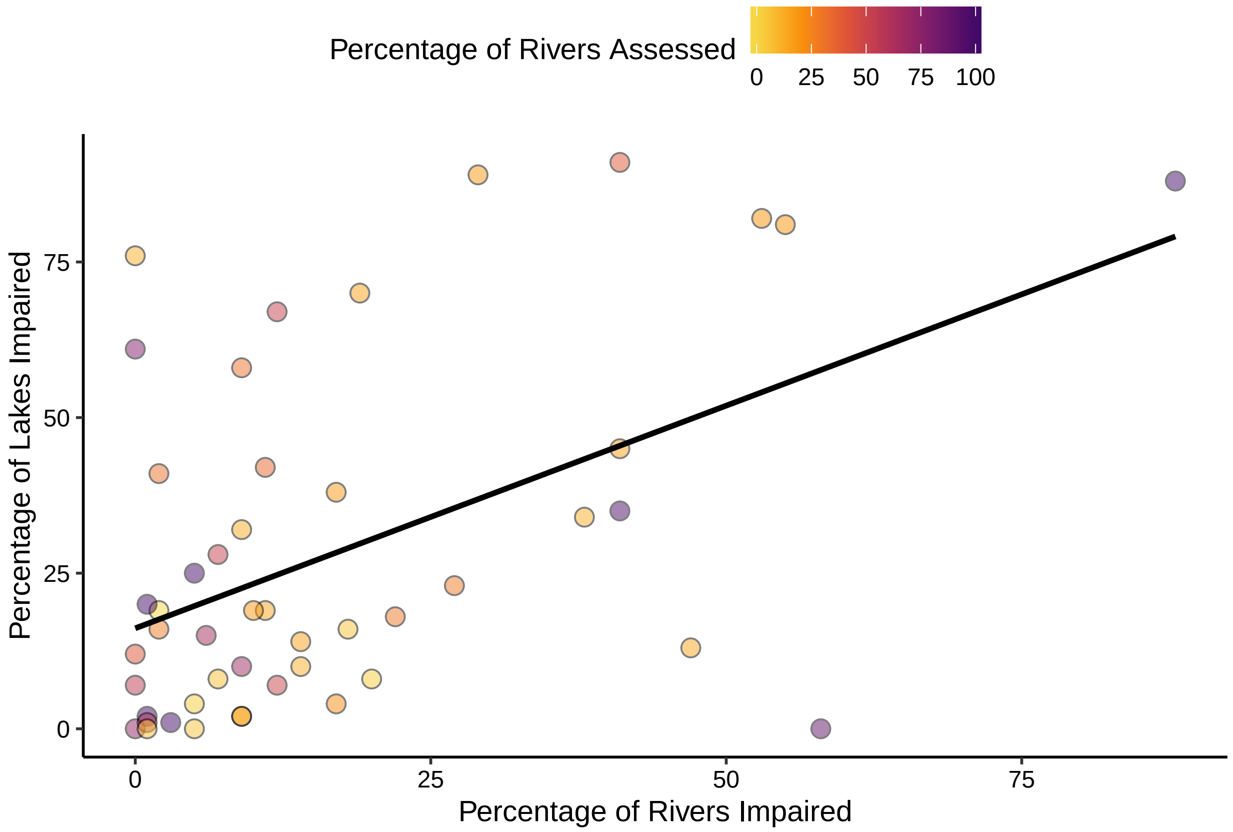
11. Create one graph that compares the data for lakes and/or rivers. This option is flexible; choose a relationship (or relationships) that seem interesting to you, and think about the implications of your findings. This graph should be edited so it follows best data visualization practices.

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

```r
ggplot(data = Rivers.and.Lakes, aes(x = Rivers.Impaired.percent, y = Lakes.Impaired.percent,
                                    fill = Rivers.Assessed.percent)) +
  geom_point(shape = 21, size = 3, alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
    scale_fill_viridis_c(option = "inferno", begin = 0.2, end = 0.9, direction = -1) +
  labs(x = "Percentage of Rivers Impaired",
       y = "Percentage of Lakes Impaired",
       fill = "Percentage of Rivers Assessed")
```

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

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



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

    There is a strong positive correlation between the percentage of impaired rivers in each state and the percentage of impaired lakes in the state. This may be attributable to the effort each state puts into assessing its rivers and lakes. However, if we look at the percentage of rivers assessed as a proxy for effort, it is still hard to detect a pattern between percentage of rivers assessed and percentage of lakes/rivers that are impaired. For further research, I would recommend running a linear regression on the percentage of lakes impaired to see if percentage of rivers impaired is in fact a good predictor. I would be sure to include an interaction term between percentage of rivers assessed and percentage of impaired rivers.