Assignment 10: Data Scraping

Pierre Mishra

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

```
# Checking working directory
getwd()
```

```
## [1] "C:/Users/Peaceful Pierre/Documents/Academics/Spring 2020/Environmental Data Analytics/Environmen
```

```
# Loading necessary libraries
library("tidyverse")
library("rvest")
```

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

```
library("ggplot2")
library("ggrepel")
```

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

```
# Setting ggplot theme
peaceful.theme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        plot.title = element_text(hjust = 0.5),
        legend.position = "right")
theme_set(peaceful.theme)
```

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.

```
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.

```
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_t

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`.

```
# 4
Rivers$Rivers.Assessed.mi2 <-
  str_replace(Rivers$Rivers.Assessed.mi2, "[,]", "")
Rivers$Rivers.Assessed.percent <-
  str_replace(Rivers$Rivers.Assessed.percent, "[%]", "")
Rivers$Rivers.Assessed.percent <-
  str_replace(Rivers$Rivers.Assessed.percent, "[*]", "")
Rivers$Rivers.Impaired.mi2 <-
  str_replace(Rivers$Rivers.Impaired.mi2, "[,]", "")
Rivers$Rivers.Impaired.percent <-
  str_replace(Rivers$Rivers.Impaired.percent, "[%]", "")
Rivers$Rivers.Impaired.percent.TMDL <-
  str_replace(Rivers$Rivers.Impaired.percent.TMDL, "[%]", "")
Rivers$Rivers.Impaired.percent.TMDL <-
  str_replace(Rivers$Rivers.Impaired.percent.TMDL, "[±]", "")

# 5
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_t

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
Lakes <- Lakes %>%
  filter(State != "Hawaii" & State != "Pennsylvania")

# 8
Lakes$Lakes.Assessed.mi2 <-          str_replace(Lakes$Lakes.Assessed.mi2, "[,]", "")
Lakes$Lakes.Assessed.percent <-      str_replace(Lakes$Lakes.Assessed.percent, "[%]", "")
Lakes$Lakes.Assessed.percent <-      str_replace(Lakes$Lakes.Assessed.percent, "[*]", "")
Lakes$Lakes.Impaired.mi2 <-          str_replace(Lakes$Lakes.Impaired.mi2, "[,]", "")
Lakes$Lakes.Impaired.percent <-      str_replace(Lakes$Lakes.Impaired.percent, "[%]", "")
Lakes$Lakes.Impaired.percent.TMDL <- str_replace(Lakes$Lakes.Impaired.percent.TMDL, "[%]", "")
Lakes$Lakes.Impaired.percent.TMDL <- str_replace(Lakes$Lakes.Impaired.percent.TMDL, "[±]", "")

# 9
Lakes$Lakes.Assessed.mi2 <-          as.numeric(Lakes$Lakes.Assessed.mi2)
```

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

```
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 NA ...
##  $ 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`.

```
rivers_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)

```
#Indiana reported that greater 100% percent of their lakes have been assessed
#which does not make sense. So first I change it to 100%.

rivers_lakes$Lakes.Assessed.percent[rivers_lakes$State == "Indiana"] <- 100

stat <- lm(data = rivers_lakes, Rivers.Assessed.percent ~ Lakes.Assessed.percent)
summary(stat)
```

```
##
## Call:
## lm(formula = Rivers.Assessed.percent ~ Lakes.Assessed.percent,
##     data = rivers_lakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.947 -19.811  -3.801  21.972  57.587
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               4.3769     9.1314   0.479 0.633980
## Lakes.Assessed.percent    0.5157     0.1278   4.035 0.000204 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.7 on 46 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.2614, Adjusted R-squared:  0.2453
## F-statistic: 16.28 on 1 and 46 DF,  p-value: 0.0002045
```

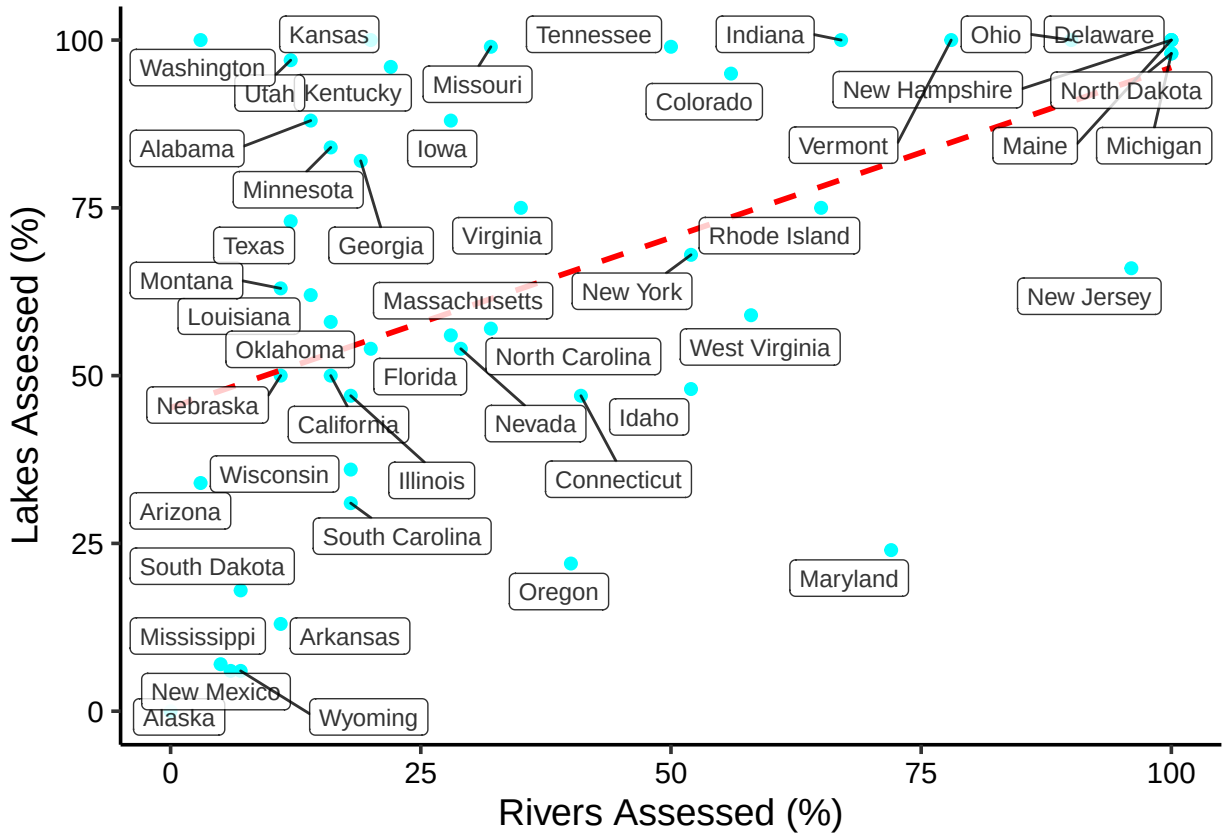
```
ggplot(rivers_lakes, aes(x = Rivers.Assessed.percent, y = Lakes.Assessed.percent)) +
  geom_point (color = "cyan", size = 1.8) +
  geom_smooth(method = 'lm', formula = y ~ x, se = FALSE,
              color = "red", lty = 2, size = 1) +
```

```
  labs (x = "Rivers Assessed (%)", y = "Lakes Assessed (%)") +
  geom_label_repel(aes(label = State), nudge_x = -2, nudge_y = -2,
                   size = 3, alpha = 0.8)
```

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

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

```
## Warning: Removed 2 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.

    Efficient management begins with extensive data collection. I wanted to see if a state with a high percentage of lake assessment is also likely to have a high percentge of river assessment or vice versa. I found that only 24.53% of variation in the percentage of lakes assessed in states were explained by the percentage of rivers assessed (linear regression, $p < 0.001$, $f(1,46) = 16.28$). From the figure, we can notice that there was a greater number of states with a higher percentage of lakes assessed than that of rivers assessed. Managers could further explore why the percentages of rivers assessed in most states is lower when compared to the percentages of lakes assessed and accordingly increase their data collection programs.