Assignment 10: Data Scraping

Yang Wang

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/26059/OneDrive/Desktop/ENV 872 R/Yang_ENV872/Assignments"
```

```
library(tidyverse)
library(rvest)

mytheme <- theme_classic(base_size = 15) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "right")
theme_set(mytheme)
```

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

```
# 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.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_text()

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`.

```r
# 4
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 = "")
Rivers$Rivers.Impaired.percent <- str_replace(Rivers$Rivers.Impaired.percent,
                                                   pattern = "([%])", replacement = "")
Rivers$Rivers.Assessed.percent <- str_replace(Rivers$Rivers.Assessed.percent,
                                                   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          : Factor w/ 50 levels "1,997","10,476",..: 3 41 20 49 29 36 17 18 2 8
##  $ Rivers.Assessed.percent     : chr  "14" "0" "3" "11" ...
##  $ Rivers.Impaired.mi          : Factor w/ 50 levels "0","1,007","1,125",..: 4 14 13 5 12 28 19 26 40
##  $ Rivers.Impaired.percent     : chr  "11" "2" "5" "14" ...
##  $ Rivers.Impaired.percent.TMDL: chr  "53" "100" "6" "2" ...
```

```r
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
```

```r
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)
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         : num  3 41 20 49 29 36 17 18 2 8 ...
##  $ Rivers.Assessed.percent    : num  14 0 3 11 16 56 41 100 20 19 ...
##  $ Rivers.Impaired.mi         : num  4 14 13 5 12 28 19 26 40 11 ...
##  $ 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.

```r
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`.

```r
# 7
Lakes <- filter(Lakes,State!="Pennsylvania"& State!="Hawaii")

# 8
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 = "")
Lakes$Lakes.Impaired.percent <- str_replace(Lakes$Lakes.Impaired.percent,
                                                 pattern = "([%])", replacement = "")
Lakes$Lakes.Assessed.percent <- str_replace(Lakes$Lakes.Assessed.percent,
                                                 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         : Factor w/ 49 levels "1,051,246","1,124,399",..: 33 37 6 43 1 14 30 2
##  $ Lakes.Assessed.percent     : chr  "88" "0" "34" "13" ...
##  $ Lakes.Impaired.mi2         : Factor w/ 47 levels "0","1,137","10,007",..: 42 2 31 39 35 4 27 20 4
##  $ 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)
```

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

```
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  33 37 6 43 1 14 30 20 2 31 ...
##  $ Lakes.Assessed.percent     : num  88 0 34 13 50 95 47 100 54 82 ...
##  $ Lakes.Impaired.mi2         : num  42 2 31 39 35 4 27 20 45 40 ...
##  $ 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`.

```
water<-full_join(Lakes, Rivers)
```

```
## Joining, 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)

```
cor.test(water$Rivers.Assessed.mi, water$Lakes.Assessed.mi2)
```

```
##
##  Pearson's product-moment correlation
##
## data:  water$Rivers.Assessed.mi and water$Lakes.Assessed.mi2
## t = 0.044967, df = 46, p-value = 0.9643
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2780285  0.2902179
## sample estimates:
##         cor
## 0.006629904
```

```r
cor.test(water$Rivers.Impaired.mi, water$Lakes.Impaired.mi2)#-0.199
```

```
##
##  Pearson's product-moment correlation
##
## data:  water$Rivers.Impaired.mi and water$Lakes.Impaired.mi2
## t = -1.3834, df = 46, p-value = 0.1732
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.45797867  0.08935891
## sample estimates:
##      cor
## -0.19985
```

```r
cor.test(water$Rivers.Assessed.percent, water$Lakes.Assessed.percent)
```

```
##
##  Pearson's product-moment correlation
##
## data:  water$Rivers.Assessed.percent and water$Lakes.Assessed.percent
## t = 3.8955, df = 43, p-value = 0.0003374
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2555024 0.6994121
## sample estimates:
##       cor
## 0.5107321
```

```r
cor.test(water$Rivers.Impaired.percent, water$Lakes.Impaired.percent)
```

```
##
##  Pearson's product-moment correlation
##
## data:  water$Rivers.Impaired.percent and water$Lakes.Impaired.percent
## t = 3.8125, df = 46, p-value = 0.0004075
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2391797 0.6795364
## sample estimates:
##       cor
## 0.4900134
```

```r
cor.test(water$Rivers.Impaired.percent.TMDL, water$Lakes.Impaired.percent.TMDL)
```

```
##
##  Pearson's product-moment correlation
##
## data:  water$Rivers.Impaired.percent.TMDL and water$Lakes.Impaired.percent.TMDL
## t = 1.1945, df = 32, p-value = 0.2411
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

```
##  -0.1414414  0.5091960
## sample estimates:
##       cor
## 0.2066065
```

```r
cor.test(water$Rivers.Assessed.mi, water$Rivers.Impaired.mi)
```

```
##
##  Pearson's product-moment correlation
##
## data:  water$Rivers.Assessed.mi and water$Rivers.Impaired.mi
## t = 0.59469, df = 48, p-value = 0.5548
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1975276  0.3554094
## sample estimates:
##        cor
## 0.08552221
```

```r
cor.test(water$Rivers.Assessed.percent, water$Rivers.Impaired.percent)
```

```
##
##  Pearson's product-moment correlation
##
## data:  water$Rivers.Assessed.percent and water$Rivers.Impaired.percent
## t = 0.074631, df = 47, p-value = 0.9408
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2711411  0.2911907
## sample estimates:
##        cor
## 0.01088547
```

```r
cor.test(water$Lakes.Assessed.mi, water$Lakes.Impaired.mi)
```

```
##
##  Pearson's product-moment correlation
##
## data:  water$Lakes.Assessed.mi and water$Lakes.Impaired.mi
## t = -0.15468, df = 46, p-value = 0.8778
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3049593  0.2630386
## sample estimates:
##         cor
## -0.02280019
```

```r
cor.test(water$Lakes.Assessed.percent, water$Lakes.Impaired.percent)#-0.156
```

```
##
##  Pearson's product-moment correlation
```

```
##
## data:  water$Lakes.Assessed.percent and water$Lakes.Impaired.percent
## t = 0.3842, df = 43, p-value = 0.7027
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2391509  0.3460811
## sample estimates:
##       cor
## 0.0584889
```

```
shapiro.test(water$Rivers.Impaired.mi)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  water$Rivers.Impaired.mi
## W = 0.95558, p-value = 0.05809
```

```
shapiro.test(water$Lakes.Impaired.mi2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  water$Lakes.Impaired.mi2
## W = 0.94965, p-value = 0.03872
```

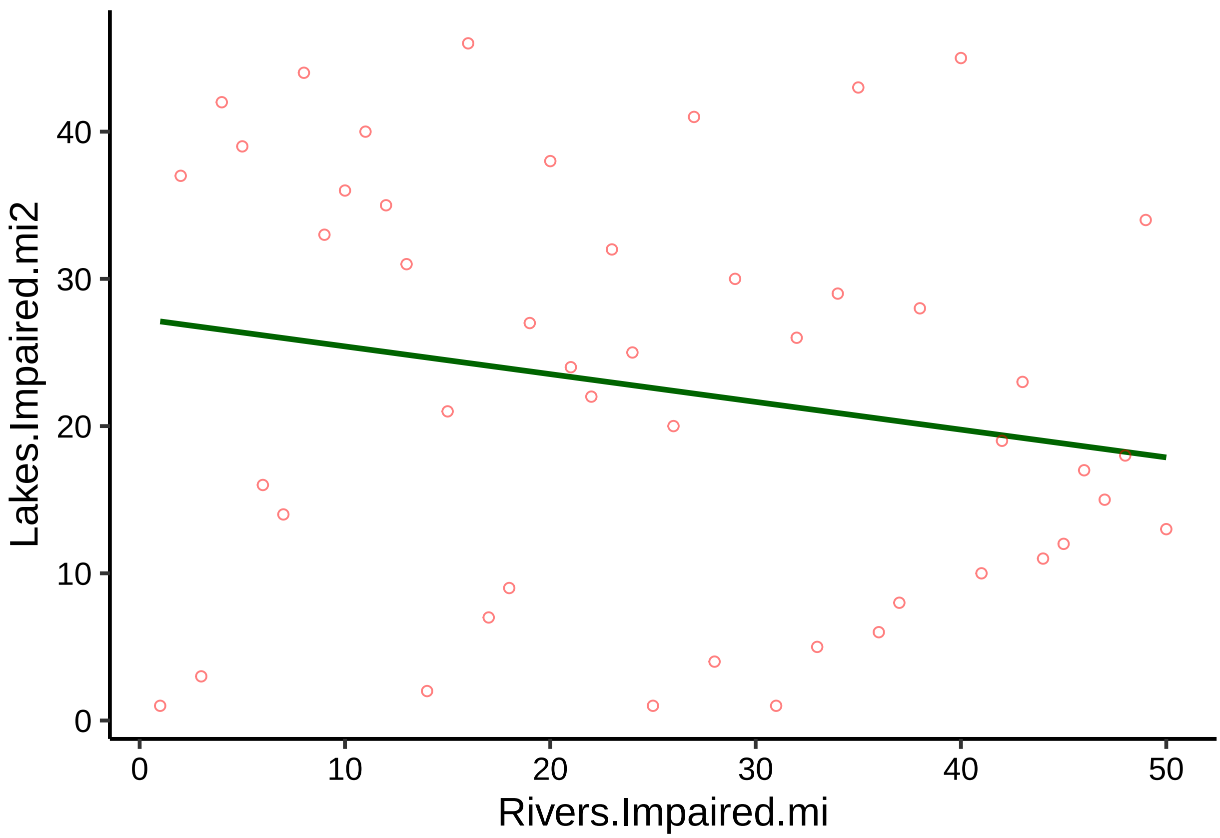
```
mo<-lm(Lakes.Impaired.percent~ Rivers.Impaired.percent + Lakes.Assessed.percent,data=water)
summary(mo)
```

```
##
## Call:
## lm(formula = Lakes.Impaired.percent ~ Rivers.Impaired.percent +
##     Lakes.Assessed.percent, data = water)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.323 -15.823  -6.263   7.916  60.312
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              14.8635     7.9555   1.868   0.0687 .
## Rivers.Impaired.percent   0.8604     0.1913   4.498 5.33e-05 ***
## Lakes.Assessed.percent    0.0165     0.1049   0.157   0.8758
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.71 on 42 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.3274, Adjusted R-squared:  0.2954
## F-statistic: 10.22 on 2 and 42 DF,  p-value: 0.0002411
```

```
plot <- ggplot(water, aes(y =Lakes.Impaired.mi2,x =Rivers.Impaired.mi))+
  geom_smooth(method="lm",se=FALSE,color="darkgreen")+
  geom_point(shape=1,alpha=0.5,color="red")+
  labs(x = "Rivers.Impaired.mi", y = "Lakes.Impaired.mi2")
print(plot)
```

```
## 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.

    My graph shows that as the miles of impaired river goes up, the square miles of impaired of lakes goes down. However the relationship is not significant(correlation test, p>0.05). It may need more data to look into this relationship.