Assignment 10: Data Scraping

Rachel Landman

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] "/Users/rmlandman/Desktop/Data Analytics/Environmental_Data_Analytics_2020/Assignments"
```

```
library(tidyverse)
library(rvest)
library(ggrepel)
library(cowplot)

# Set theme
mytheme <- theme_bw(base_size = 12) +
  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-nutr 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,
                                          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.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.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$Lakes.Assessed.acres <- str_replace(Lakes$Lakes.Assessed.acres ,
                                          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  "430.976" "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)
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       : num  431 5981 114976 64778 NA ...
##  $ Lakes.Assessed.percent     : num  88 0 34 13 50 95 47 100 54 82 ...
##  $ Lakes.Impaired.acres       : 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`.

```
EPA.waterquality <- full_join(Rivers, Lakes)
```

```
## 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)

```
#### Is there a correlation between Lakes and River
#### assessments, impairments, and TMDLs within each state?

### Impaired

lakes.rivers.impaired.regression <- lm(EPA.waterquality$Lakes.Impaired.percent ~ EPA.waterquality$River
summary(lakes.rivers.impaired.regression)
```

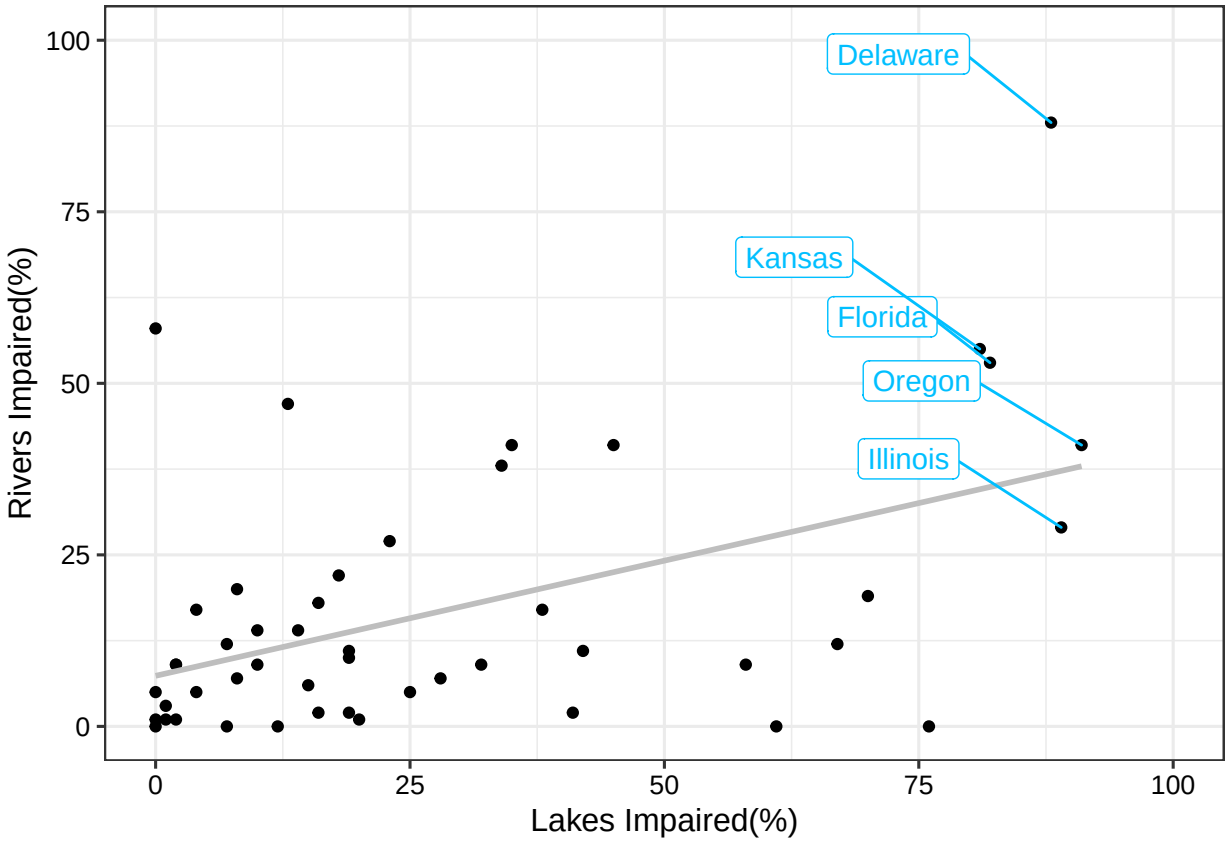
```
##
## Call:
## lm(formula = EPA.waterquality$Lakes.Impaired.percent ~ EPA.waterquality$Rivers.Impaired.percent)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.641 -15.931  -7.290   9.486  59.856
##
```

```
## Coefficients:
##                                       Estimate Std. Error t value
## (Intercept)                            16.1439     4.7703   3.384
## EPA.waterquality$Rivers.Impaired.percent   0.7155     0.1877   3.813
##                                       Pr(>|t|)
## (Intercept)                           0.001468 **
## EPA.waterquality$Rivers.Impaired.percent 0.000407 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.76 on 46 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.2401, Adjusted R-squared:  0.2236
## F-statistic: 14.54 on 1 and 46 DF,  p-value: 0.0004075
```

```
Impaired <- ggplot(EPA.waterquality, aes (x=Lakes.Impaired.percent, y=Rivers.Impaired.percent))+
  geom_point()+
  ylim(0, 100)+
  xlim(0,100)+
   geom_smooth(method = "lm", se = FALSE, color = "gray")+
   labs(x = "Lakes Impaired(%)",
        y = "Rivers Impaired(%)")+
   geom_label_repel(data = subset(EPA.waterquality, State %in% c("Kansas", "Florida", "Delaware", "Illi
                    aes(label = State), color = "deepskyblue1", nudge_x = -15, nudge_y = 10)
print(Impaired)
```

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

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




```
#### Assessed

lakes.rivers.assessed.regression <- lm(EPA.waterquality$Lakes.Assessed.percent ~
                                       EPA.waterquality$Rivers.Assessed.percent)
summary(lakes.rivers.assessed.regression)
```

```
##
## Call:
## lm(formula = EPA.waterquality$Lakes.Assessed.percent ~ EPA.waterquality$Rivers.Assessed.percent)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.327 -19.275  -1.591  19.891  80.398
##
## Coefficients:
##                                       Estimate Std. Error t value
## (Intercept)                            45.0789     6.6900   6.738
## EPA.waterquality$Rivers.Assessed.percent   0.5451     0.1365   3.993
##                                       Pr(>|t|)
## (Intercept)                           2.25e-08 ***
## EPA.waterquality$Rivers.Assessed.percent 0.000233 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.84 on 46 degrees of freedom
##   (2 observations deleted due to missingness)
```

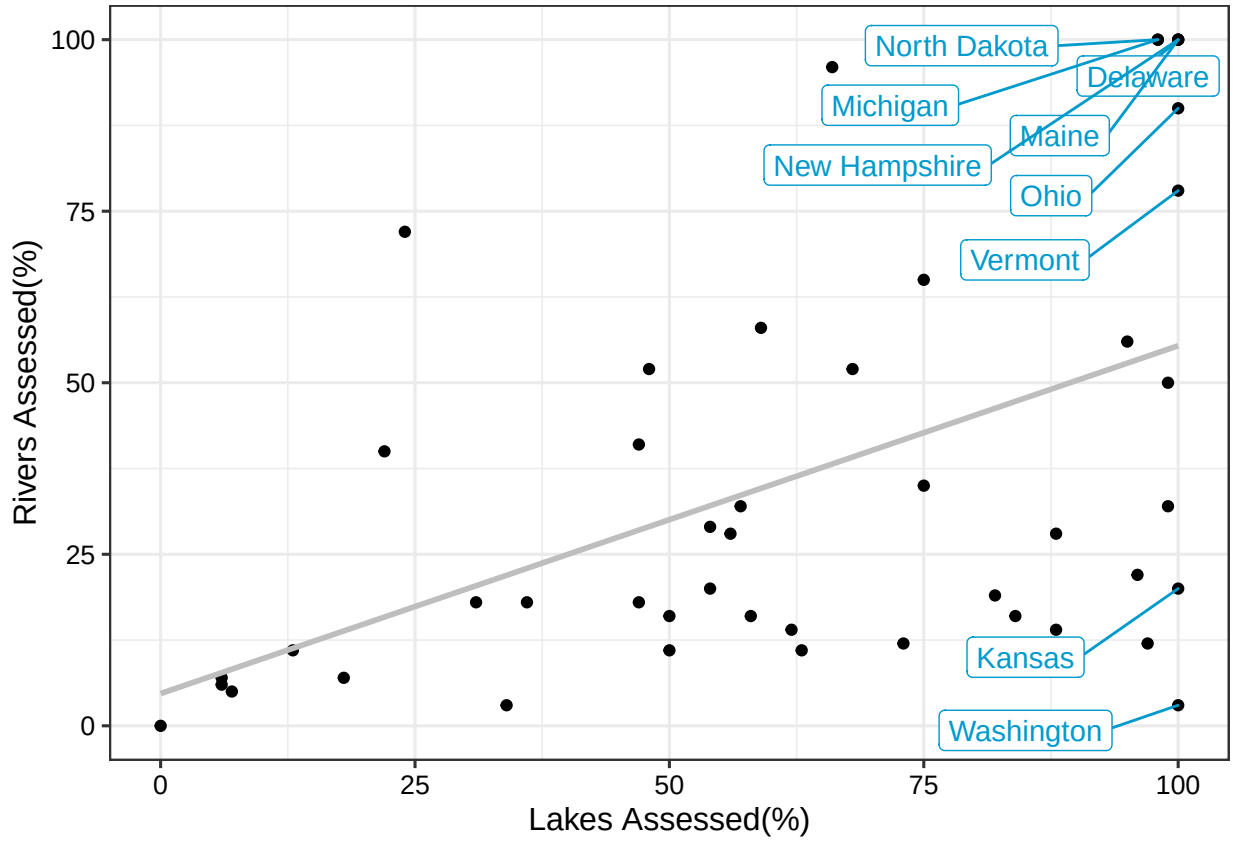
```
## Multiple R-squared:  0.2574, Adjusted R-squared:  0.2413
## F-statistic: 15.94 on 1 and 46 DF,  p-value: 0.0002329
```

```
Assessed <- ggplot(EPA.waterquality, aes (x=Lakes.Assessed.percent, y=Rivers.Assessed.percent))+
  geom_point()+
  ylim(0, 100)+
  xlim(0,100)+
   geom_smooth(method = "lm", se = FALSE, color = "gray")+
   labs(x = "Lakes Assessed(%)",
        y = "Rivers Assessed(%)")+
  geom_label_repel(data = subset(EPA.waterquality, State %in% c("Delaware", "New Hampshire", "Michigan"
                   aes(label = State), color = "deepskyblue3", nudge_x = -15, nudge_y = -10)
print(Assessed)
```

```
## Warning: Removed 3 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```

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



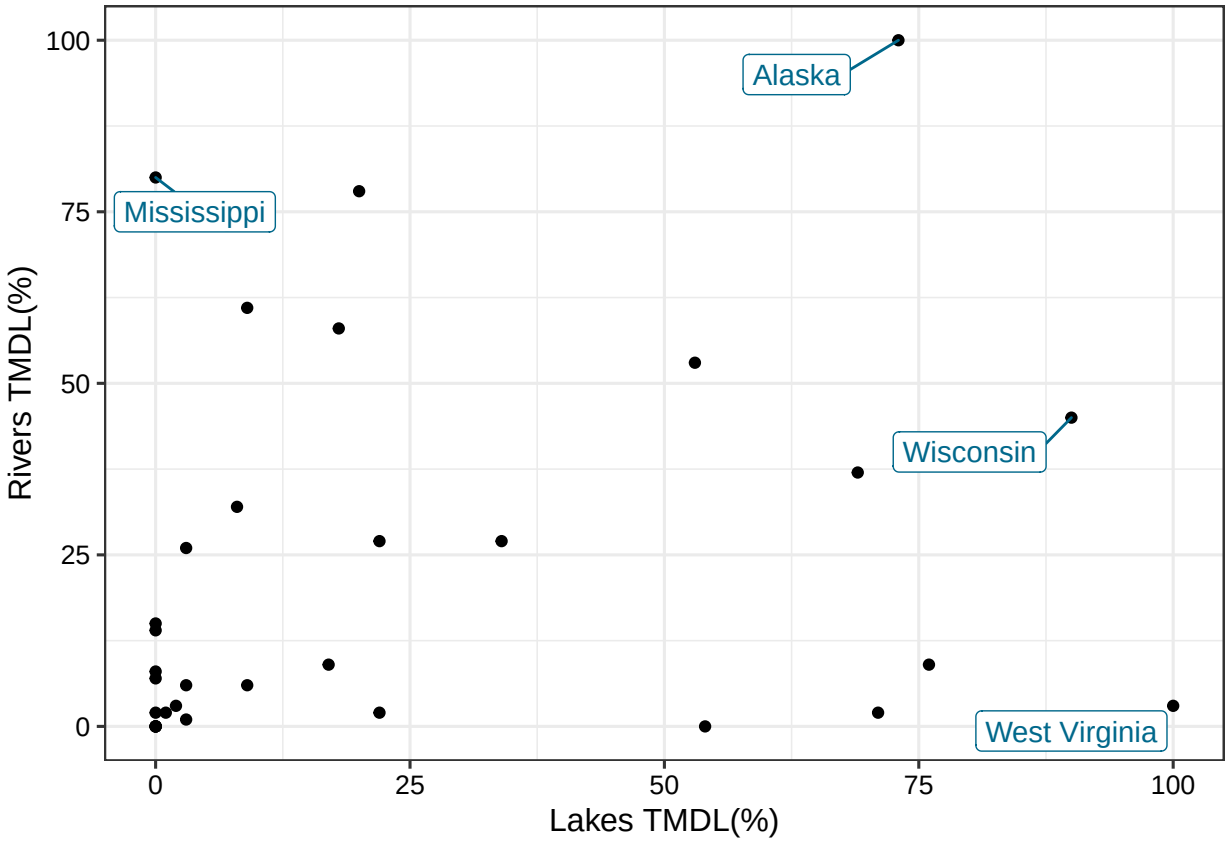

```
### TMDL

lakes.rivers.TMDL.regression <- lm(EPA.waterquality$Lakes.Impaired.percent.TMDL ~ EPA.waterquality$Rive
summary(lakes.rivers.TMDL.regression)
```

```
##
## Call:
## lm(formula = EPA.waterquality$Lakes.Impaired.percent.TMDL ~ EPA.waterquality$Rivers.Impaired.percent
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.090 -17.675 -15.840   9.055  81.993
##
## Coefficients:
##                                             Estimate Std. Error t value
## (Intercept)                                  17.3417     6.7616   2.565
## EPA.waterquality$Rivers.Impaired.percent.TMDL   0.2219     0.1857   1.195
##                                             Pr(>|t|)
## (Intercept)                                   0.0152 *
## EPA.waterquality$Rivers.Impaired.percent.TMDL   0.2411
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.46 on 32 degrees of freedom
##   (16 observations deleted due to missingness)
## Multiple R-squared:  0.04269,    Adjusted R-squared:  0.01277
## F-statistic: 1.427 on 1 and 32 DF,  p-value: 0.2411
```

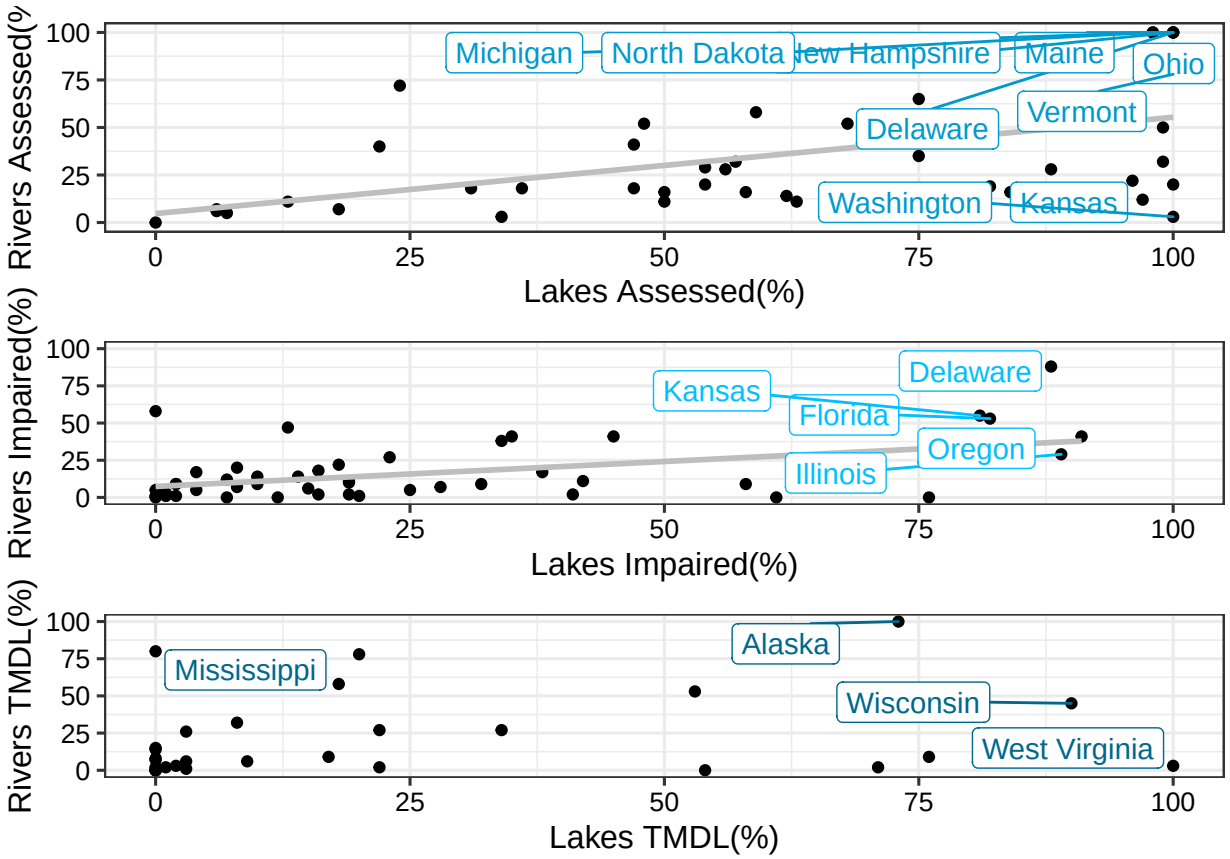
```
TMDL <- ggplot(EPA.waterquality, aes (x=Lakes.Impaired.percent.TMDL, y=Rivers.Impaired.percent.TMDL))+
  geom_point()+
   labs(x = "Lakes TMDL(%)",
        y = "Rivers TMDL(%)")+
  geom_label_repel(data = subset(EPA.waterquality, State %in% c("West Virginia", "Wisconsin", "Alaska",
                   aes(label = State), color = "deepskyblue4", nudge_x = -10, nudge_y = -5)
print(TMDL)
```

```
## Warning: Removed 16 rows containing missing values (geom_point).
```




```
plot_grid(Assessed,Impaired,TMDL, nrow = 3, align = 'h', rel_heights = c(1.75, 1.5,1.5))
```

```
## Warning: Removed 3 rows containing non-finite values (stat_smooth).
## Warning: Removed 3 rows containing missing values (geom_point).
## Warning: Removed 2 rows containing missing values (geom_label_repel).
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
## Warning: Removed 2 rows containing missing values (geom_point).
## Warning: Removed 16 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 significant relationship between the percentage of lakes and rivers that are assessed in each state (linear regression, Adj R2 = 0.2413, df = 46, p-value < 0.0003) and also a significant relationship between the percentage of lakes and rivers that are impaired in each state (linear regression, Adj R2 = 0.2236, df = 46, p-value < 0.0005). There is not a significant relationship between the percentage of impaired lakes and rivers that have a TMDL in each state. It seems like some states focus more on regulating maximum daily loads in rivers and others focus more on lakes. The labels on the graphs highlight states that have high percentages of lakes or rivers that are assessed, impaired, or have TMDLs. It is interesting to see that few states show up on all three graphs. Therefore states with a high percentage of rivers/lakes assessed don't necessarily have a high percentage of rivers/lakes impaired or a high percentage of TMDLs. It would be interesting to examine the location of the rivers and lake with TMDLs in comparison to those without. Specifically, you could do spatial analysis about location in urban vs. rural areas.