Assignment 5: Water Quality in Lakes

Jake Greif

OVERVIEW

This exercise accompanies the lessons in Hydrologic Data Analysis on water quality in lakes

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 HTML file.
5. After Knitting, submit the completed exercise (HTML file) to the dropbox in Sakai. Add your last name into the file name (e.g., "A05_Salk.html") prior to submission.

The completed exercise is due on 2 October 2019 at 9:00 am.

Setup

1. Verify your working directory is set to the R project file,
2. Load the tidyverse, lubridate, and LAGOSNE packages.
3. Set your ggplot theme (can be theme_classic or something else)
4. Load the LAGOSdata database and the trophic state index csv file we created on 2019/09/27.

```
getwd()
```

```
## [1] "/Users/jakegreif/Duke/Fall_2019/Hydrologic_Data_Analysis"
```

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------- tidyverse 1.
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------- tidyverse_conflicts
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(LAGOSNE)

theme_set(theme_classic())
```

```r
options(scipen = 100)

# Load LAGOSNE data into R session
LAGOSdata <- lagosne_load()
```

```
## Warning in `_f`(version = version, fpath = fpath): LAGOSNE version
## unspecified, loading version: 1.087.3
```

```r
# Load TSI csv
LAGOStrophic <- read_csv("./Data/LAGOStrophic.csv")
```

```
## Parsed with column specification:
## cols(
##   lagoslakeid = col_double(),
##   sampledate = col_date(format = ""),
##   chla = col_double(),
##   tp = col_double(),
##   secchi = col_double(),
##   gnis_name = col_character(),
##   lake_area_ha = col_double(),
##   state = col_character(),
##   state_name = col_character(),
##   sampleyear = col_double(),
##   samplemonth = col_double(),
##   season = col_double(),
##   TSI.chl = col_double(),
##   TSI.secchi = col_double(),
##   TSI.tp = col_double(),
##   trophic.class = col_character()
## )
```

Trophic State Index

5. Similar to the trophic.class column we created in class (determined from TSI.chl values), create two additional columns in the data frame that determine trophic class from TSI.secchi and TSI.tp (call these trophic.class.secchi and trophic.class.tp).

```r
# Add secchi TSI column
LAGOStrophic <-
  mutate(LAGOStrophic, trophic.class.secchi =
           ifelse(TSI.secchi < 40, "Oligotrophic",
                  ifelse(TSI.secchi < 50, "Mesotrophic",
                         ifelse(TSI.secchi < 70, "Eutrophic",
                                "Hypereutrophic"))))

# Add TP TSI column
LAGOStrophic <-
  mutate(LAGOStrophic, trophic.class.tp =
           ifelse(TSI.tp < 40, "Oligotrophic",
                  ifelse(TSI.tp < 50, "Mesotrophic",
                         ifelse(TSI.tp < 70, "Eutrophic",
                                "Hypereutrophic"))))

# Convert to factor
```

```
LAGOStrophic$trophic.class.secchi <-
  factor(LAGOStrophic$trophic.class.secchi,
         levels = c("Oligotrophic", "Mesotrophic",
                    "Eutrophic", "Hypereutrophic"))

LAGOStrophic$trophic.class.tp <-
  factor(LAGOStrophic$trophic.class.tp,
         levels = c("Oligotrophic", "Mesotrophic",
                    "Eutrophic", "Hypereutrophic"))

LAGOStrophic$trophic.class <-
  factor(LAGOStrophic$trophic.class,
         levels = c("Oligotrophic", "Mesotrophic",
                    "Eutrophic", "Hypereutrophic"))
```

6. How many observations fall into the four trophic state categories for the three metrics (trophic.class, trophic.class.secchi, trophic.class.tp)? Hint: `count` function.

```
count(LAGOStrophic, vars = trophic.class)
```

```
## # A tibble: 4 x 2
##   vars               n
##   <fct>          <int>
## 1 Oligotrophic    3298
## 2 Mesotrophic    15413
## 3 Eutrophic      41861
## 4 Hypereutrophic 14379
```

```
count(LAGOStrophic, vars = trophic.class.secchi)
```

```
## # A tibble: 4 x 2
##   vars               n
##   <fct>          <int>
## 1 Oligotrophic   16110
## 2 Mesotrophic    25083
## 3 Eutrophic      28659
## 4 Hypereutrophic  5099
```

```
count(LAGOStrophic, vars = trophic.class.tp)
```

```
## # A tibble: 4 x 2
##   vars               n
##   <fct>          <int>
## 1 Oligotrophic   19861
## 2 Mesotrophic    23023
## 3 Eutrophic      24839
## 4 Hypereutrophic  7228
```

7. Find and quantify instances where trophic.class and trophic.class.secchi differ. Do the same comparing trophic.class and trophic.class.tp.

```
# Add tc.tci column
LAGOStrophic <-
  mutate(LAGOStrophic, tc.tci =
           ifelse(trophic.class == trophic.class.secchi,
                  "True", "False"))
```

```r
# Add tc.tcp column
LAGOStrophic <-
  mutate(LAGOStrophic, tc.tcp =
           ifelse(trophic.class == trophic.class.tp,
                  "True", "False"))

# Get count of TSI columns matching/differing
count(LAGOStrophic, vars = tc.tci)
```

```
## # A tibble: 2 x 2
##   vars      n
##   <chr> <int>
## 1 False 47433
## 2 True  27518
```

```r
47433/74951
```

```
## [1] 0.6328535
```

```r
count(LAGOStrophic, vars = tc.tcp)
```

```
## # A tibble: 2 x 2
##   vars      n
##   <chr> <int>
## 1 False 48374
## 2 True  26577
```

```r
48374/74951
```

```
## [1] 0.6454083
```

What proportion of observations do these metrics differ?

> Trophic.class differs from trophic.class.secchi in 63.3% of the lakes, and it differs from trophic.class.tp in 64.5% of the lakes.

Note: To take this further, a researcher might determine which trophic classes are susceptible to being differently categorized by the different metrics and whether certain metrics are prone to categorizing trophic class as more or less eutrophic. This would entail more complex code.

Nutrient Concentrations

8. Create a data frame that includes the columns lagoslakeid, sampledate, tn, tp, state, and state_name. Mutate this data frame to include sampleyear and samplemonth columns as well. Call this data frame LAGOSNandP.

```r
# Exploring the data types that are available
LAGOSlocus <- LAGOSdata$locus
LAGOSstate <- LAGOSdata$state
LAGOSnutrient <- LAGOSdata$epi_nutr

# Tell R to treat lakeid as a factor, not a numeric value
LAGOSlocus$lagoslakeid <- as.factor(LAGOSlocus$lagoslakeid)
LAGOSnutrient$lagoslakeid <- as.factor(LAGOSnutrient$lagoslakeid)

# Join data frames
LAGOSlocations <- left_join(LAGOSlocus, LAGOSstate, by = "state_zoneid")
```

```
# Create LAGOSNandP data frame
LAGOSNandP <- left_join(LAGOSnutrient, LAGOSlocations,
                        by = "lagoslakeid") %>%
  select("lagoslakeid", "sampledate", "tn",
           "tp", "state", "state_name") %>%
    mutate(sampleyear = year(sampledate),
         samplemonth = month(sampledate))
```

```
## Warning: Column `lagoslakeid` joining factors with different levels,
## coercing to character vector
```

```
class(LAGOSNandP$samplemonth)
```

```
## [1] "numeric"
```

```
LAGOSNandP$samplemonth <- as.factor(LAGOSNandP$samplemonth)
LAGOSNandP$sampleyear <- as.factor(LAGOSNandP$sampleyear)
```

9. Create two violin plots comparing TN and TP concentrations across states. Include a 50th percentile line inside the violins.

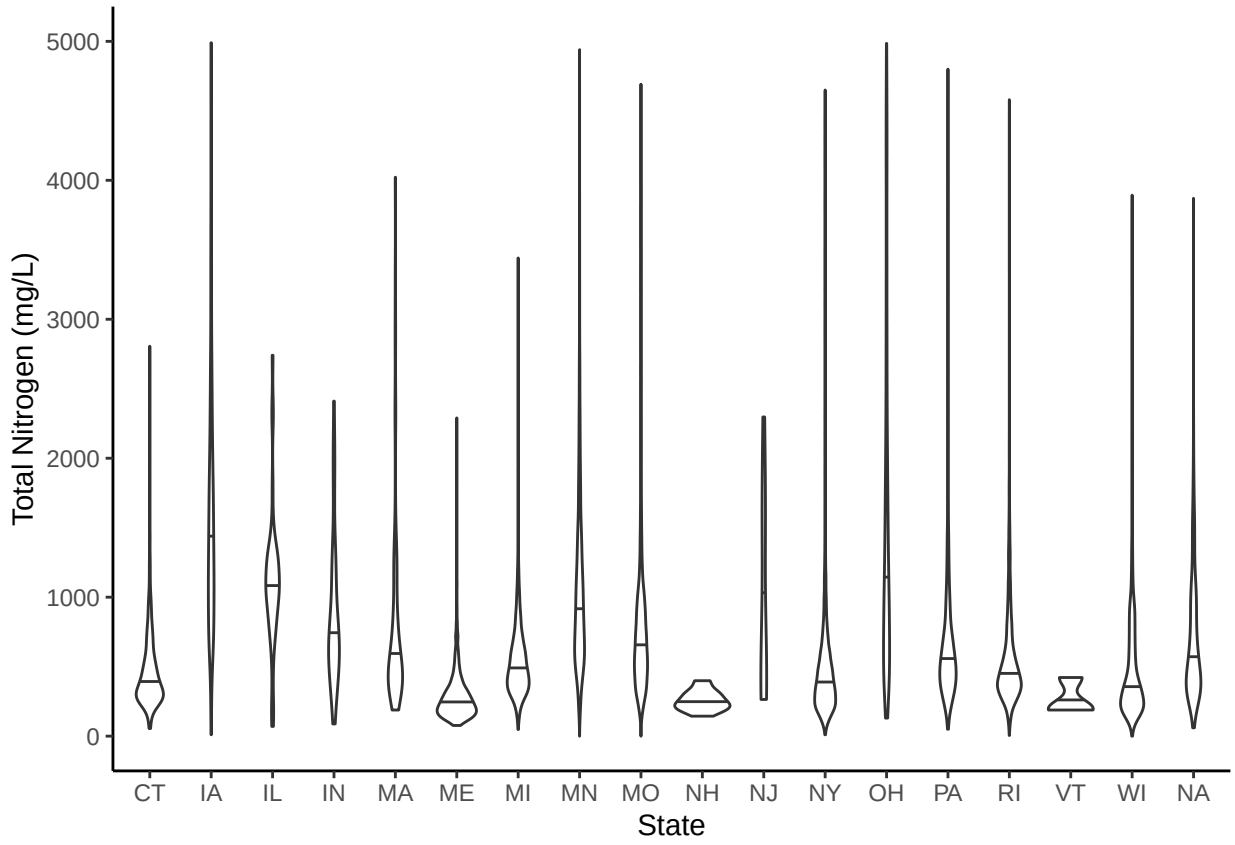
```
# TN violin graph
stateTNviolin <- ggplot(LAGOSNandP, aes(x = state, y = tn)) +
  labs(x = "State", y = "Total Nitrogen (mg/L)") +
  scale_y_continuous(limits = c(0,5000)) +
  geom_violin(draw_quantiles = 0.50)
print(stateTNviolin)
```

```
## Warning: Removed 774906 rows containing non-finite values (stat_ydensity).
```

```
## Warning in regularize.values(x, y, ties, missing(ties)): collapsing to
## unique 'x' values

## Warning in regularize.values(x, y, ties, missing(ties)): collapsing to
## unique 'x' values

## Warning in regularize.values(x, y, ties, missing(ties)): collapsing to
## unique 'x' values
```

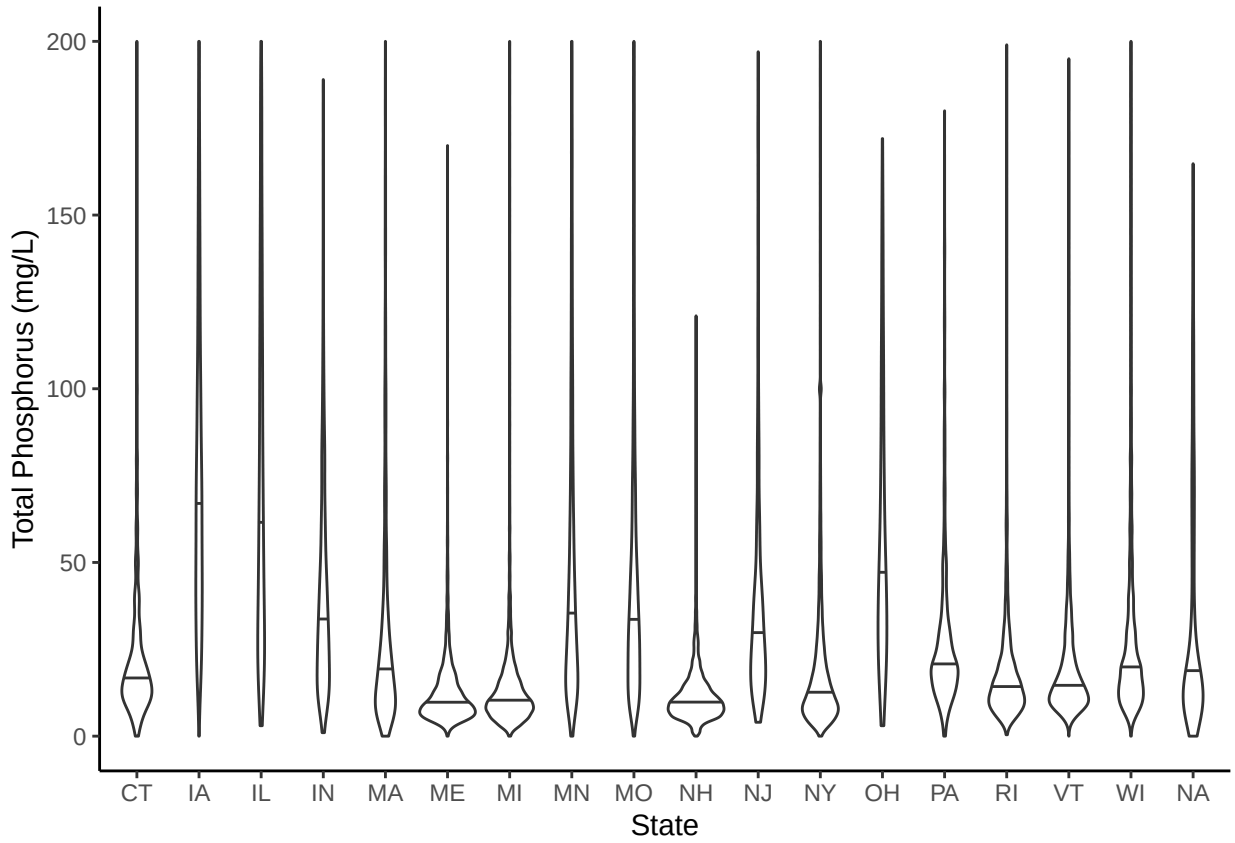



```
# TP violin graph
stateTPviolin <- ggplot(LAGOSNandP, aes(x = state, y = tp)) +
  labs(x = "State", y = "Total Phosphorus (mg/L)") +
  scale_y_continuous(limits = c(0,200)) +
  geom_violin(draw_quantiles = 0.50)
print(stateTPviolin)
```

```
## Warning: Removed 676460 rows containing non-finite values (stat_ydensity).

## Warning: collapsing to unique 'x' values

## Warning: collapsing to unique 'x' values
```



Which states have the highest and lowest median concentrations?

TN: Iowa has the highest, and Maine has the lowest.

TP: Iowa has the highest, and Maine has the lowest.

Which states have the highest and lowest concentration ranges?

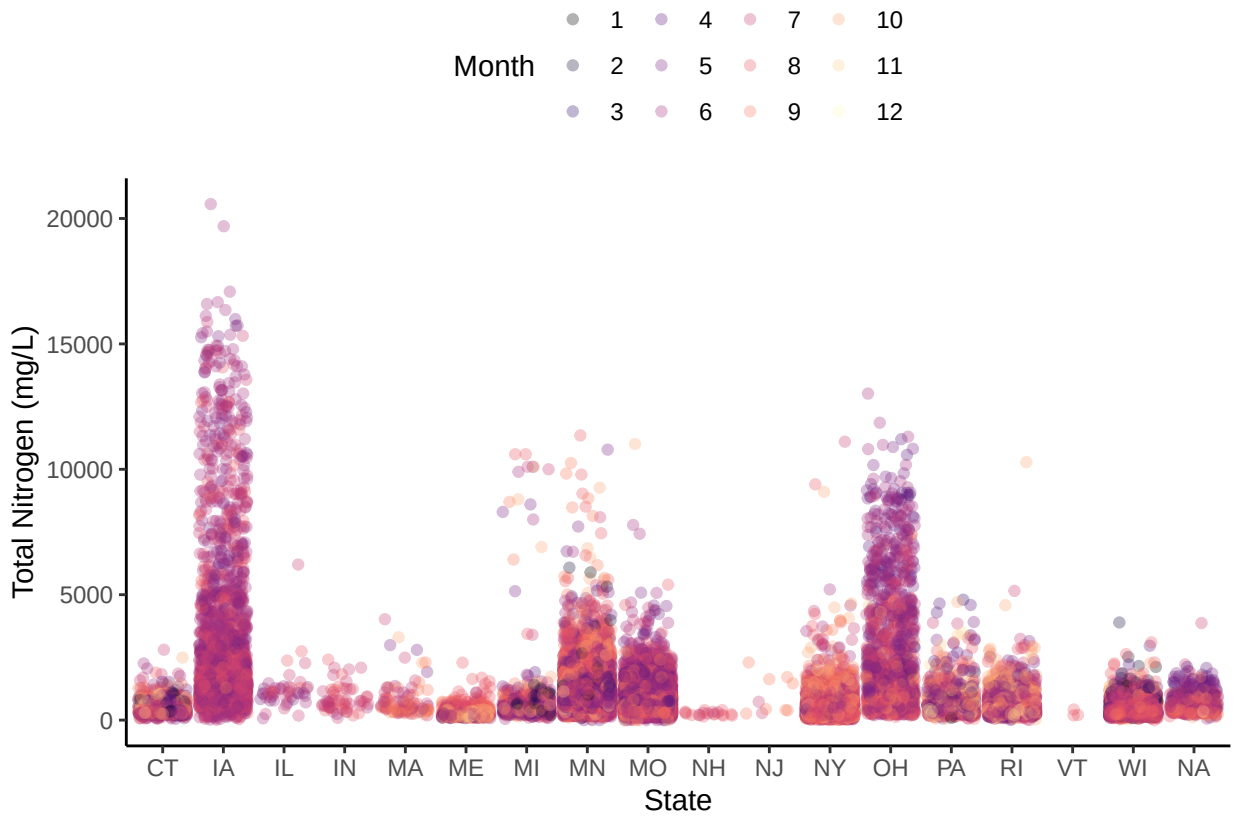
TN: Iowa has the highest, and Vermont has the lowest.

TP: Illinois has the highest, and Pennsylvania has the lowest.

10. Create two jitter plots comparing TN and TP concentrations across states, with samplemonth as the color. Choose a color palette other than the ggplot default.

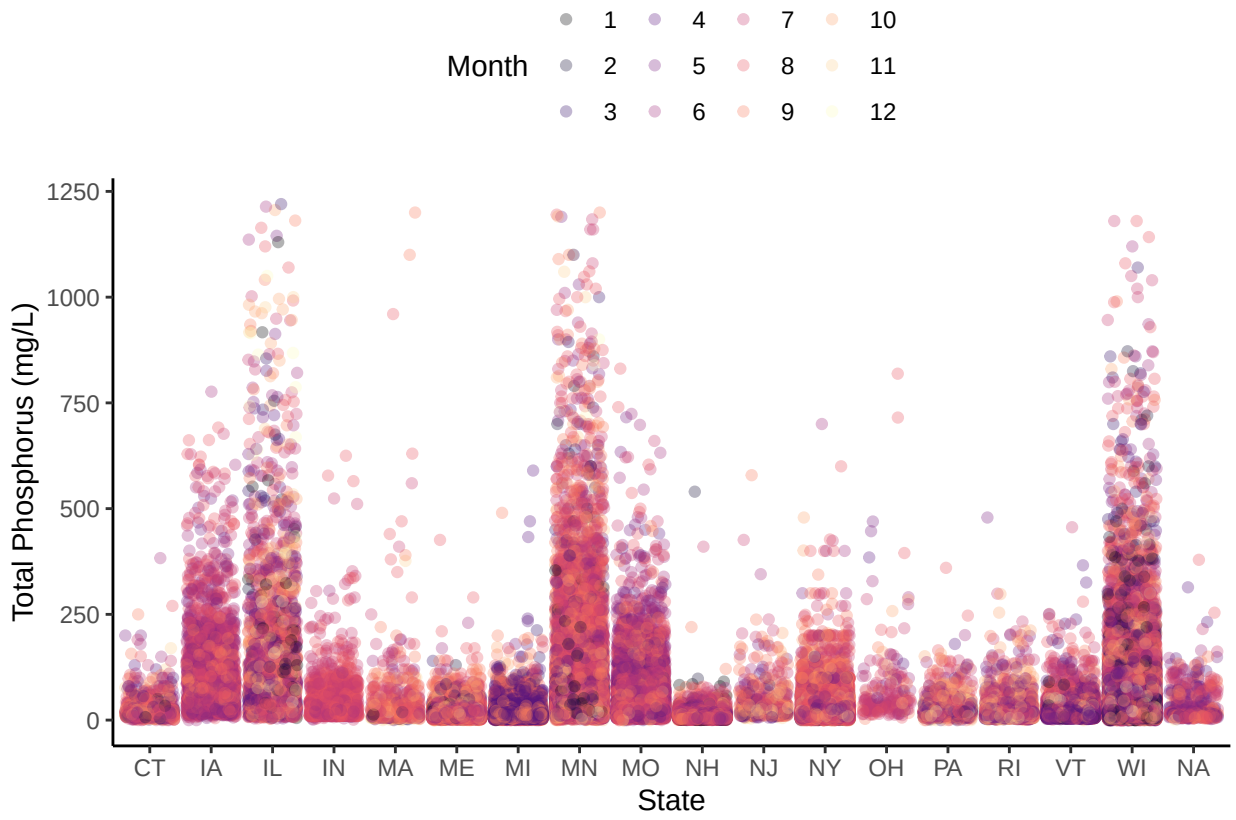
```
# TN jitter plot
stateTNjitter <- ggplot(LAGOSNandP, aes(x = state, y = tn, color = samplemonth)) +
  geom_jitter(alpha = 0.3) +
  labs(x = "State", y = "Total Nitrogen (mg/L)", color = "Month") +
  theme(legend.position = "top") +
  scale_color_viridis_d(option = "magma")
print(stateTNjitter)
```

```
## Warning: Removed 774226 rows containing missing values (geom_point).
```




```
# TP jitter plot
stateTPjitter <- ggplot(LAGOSNandP, aes(x = state, y = tp, color = samplemonth)) +
  geom_jitter(alpha = 0.3) +
  labs(x = "State", y = "Total Phosphorus (mg/L)", color = "Month") +
  theme(legend.position = "top") +
  scale_color_viridis_d(option = "magma")
print(stateTPjitter)
```

```
## Warning: Removed 672861 rows containing missing values (geom_point).
```




```
# Find counts of measurements by state
LAGOSN <- select(LAGOSNandP, -"tp") %>%
  na.omit()

count(LAGOSN, var = state)
```

```
## # A tibble: 17 x 2
##    var       n
##    <chr> <int>
##  1 CT      916
##  2 IA     2649
##  3 IL       46
##  4 IN       57
##  5 MA       95
##  6 ME      762
##  7 MI      885
##  8 MN     8604
##  9 MO    11503
## 10 NH       19
## 11 NJ       10
## 12 NY     8091
## 13 OH     1502
## 14 PA     1044
## 15 RI     2836
## 16 VT        3
## 17 WI     2416
```

```
TN.count.by.state <- count(LAGOSN, var = samplemonth, state)

LAGOSP <- select(LAGOSNandP, -"tn") %>%
  na.omit()

count(LAGOSP, var = state)
```

```
## # A tibble: 17 x 2
##    var       n
##    <chr> <int>
##  1 CT     1222
##  2 IA     2920
##  3 IL     2632
##  4 IN     1340
##  5 MA      657
##  6 ME    11987
##  7 MI    10250
##  8 MN    11186
##  9 MO    11786
## 10 NH     8164
## 11 NJ      516
## 12 NY    21343
## 13 OH      175
## 14 PA     1240
## 15 RI     3612
## 16 VT     7980
## 17 WI    45743
```

```
TP.count.by.state <- count(LAGOSP, var = samplemonth, state)
```

Which states have the most samples? How might this have impacted total ranges from #9?

> TN: Missouri has the most TN samples.
>
> TP: Wisconsin has the most TP samples.
>
> This likely impacted the total ranges from #9 because the more samples that are taken, the more likely it is that some will be outliers. Moreover, TN and TP concentrations likely flucuate from season to season, so if the samples were taken throughout the year the range is likely to be greater, versus collecting samples during only a few months.

Which months are sampled most extensively? Does this differ among states?

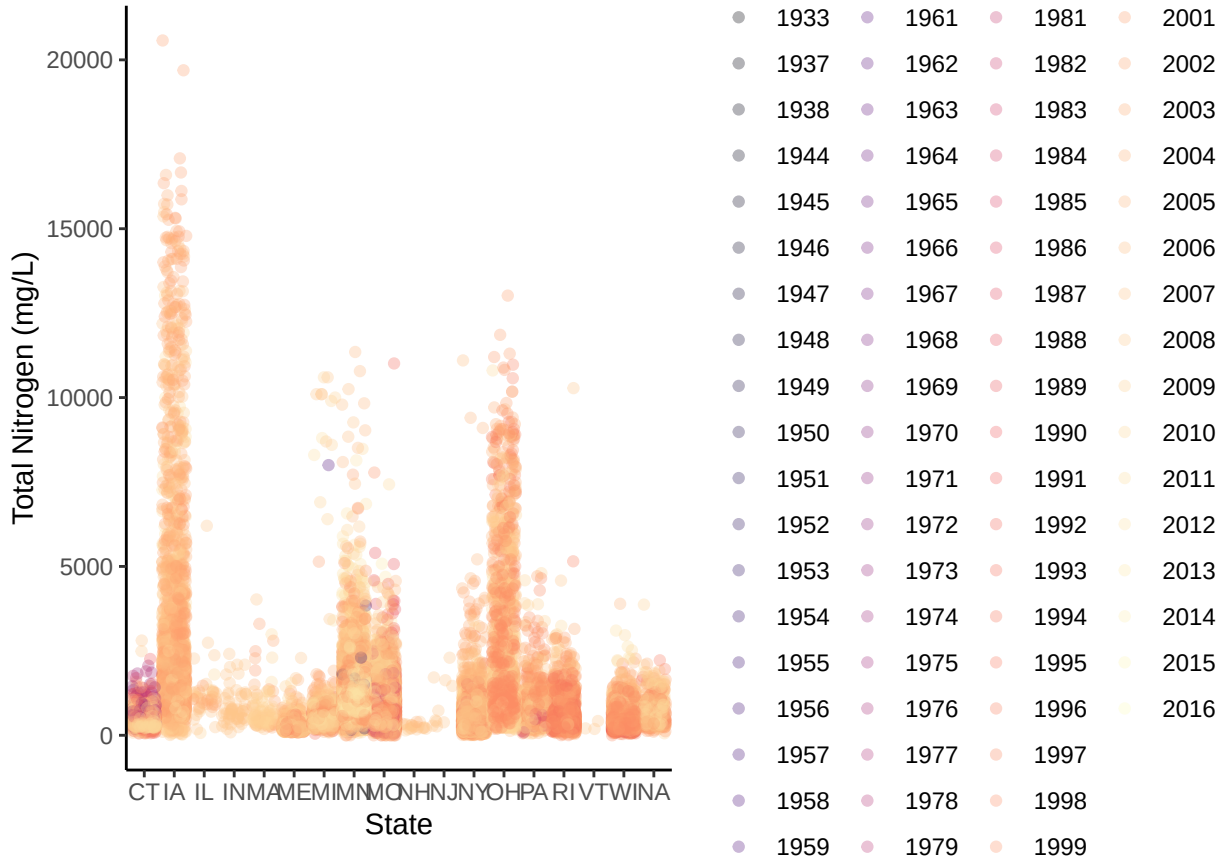
> TN: July and August are sampled most extensively.
>
> TP: July and August are sampled most extensively.
>
> Some states have consistent sampling throughout the summer (May-August), but generally all states sample the most in July and August.

11. Create two jitter plots comparing TN and TP concentrations across states, with sampleyear as the color. Choose a color palette other than the ggplot default.

```
# TN jitter plot
stateTNjitter.yr <- ggplot(LAGOSNandP, aes(x = state, y = tn, color = sampleyear)) +
  geom_jitter(alpha = 0.3) +
  labs(x = "State", y = "Total Nitrogen (mg/L)", color = "Year") +
  theme(legend.position = "right") +
```

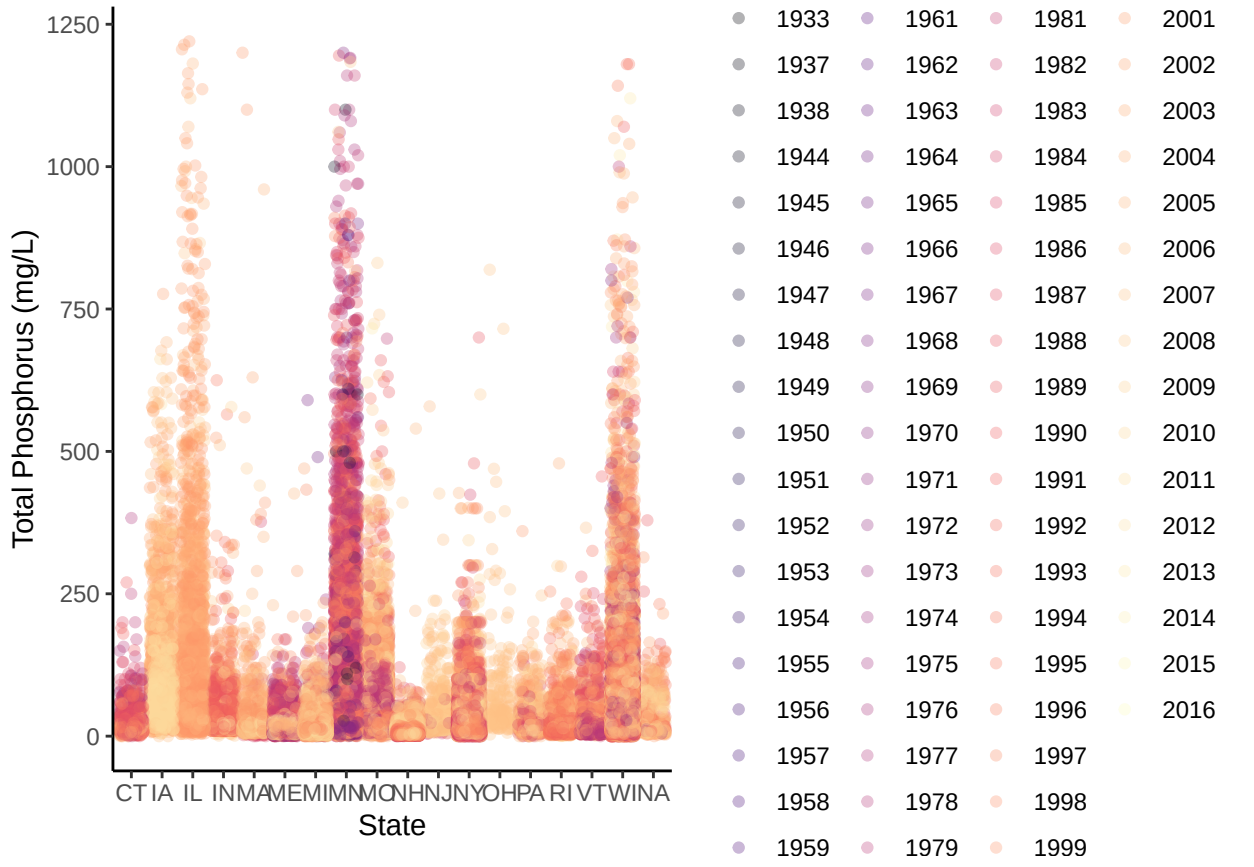
```
  scale_color_viridis_d(option = "magma")
print(stateTNjitter.yr)
```

```
## Warning: Removed 774226 rows containing missing values (geom_point).
```




```
# TP jitter plot
stateTPjitter.yr <- ggplot(LAGOSNandP, aes(x = state, y = tp, color = sampleyear)) +
  geom_jitter(alpha = 0.3) +
  labs(x = "State", y = "Total Phosphorus (mg/L)", color = "Year") +
  theme(legend.position = "right") +
  scale_color_viridis_d(option = "magma")
print(stateTPjitter.yr)
```

```
## Warning: Removed 672861 rows containing missing values (geom_point).
```




```
TN.count.year <- count(LAGOSN, var = sampleyear)
TP.count.year <- count(LAGOSP, var = sampleyear)

TN.year.state <- count(LAGOSN, var = sampleyear, state)
TP.year.state <- count(LAGOSP, var = sampleyear, state)
```

Which years are sampled most extensively? Does this differ among states?

TN: 2009

TP: 2009

The majority of states performed their most extensive sampling in the second half of the 2000s, primarily 2006-2009. However, the most extensive TP sampling years were not consistent, ranging from 1976 to 2010. For TP, Minnesota did their most extensive long-term sampling in the 1980s.

Reflection

12. What are 2-3 conclusions or summary points about lake water quality you learned through your analysis?

    There are a variety of ways to measure lake water quality, but they are not perfect. Therefore, it's important to use multiple measures to get the most complete picture of lake quality.

13. What data, visualizations, and/or models supported your conclusions from 12?

    All of lesson 9, and looking at the data visualized as violin, jitter, and bar graphs. The comparison of the TSI values made the conclusions mentioned above clear and easy to comprehend.

14. Did hands-on data analysis impact your learning about water quality relative to a theory-based lesson? If so, how?

    Yes, but I think that the TSI portion of the lessons was based in some theory, so I feel like there was a lot of overlap between hands-on data analysis and the theory behind our analysis. Being able to manipulate the data myself allowed me to get familiar with it at my own pace, and I was able to think about why the data is the way it is by comparing it to my prior theory-based lessons.

15. How did the real-world data compare with your expectations from theory?

    As I mentioned in #14, I felt that there was a lot of overlap for this section, so this is a difficult question to answer. I know that quantifying lake quality is not a perfect science, so my understanding of the "theory" is that it is difficult to generalize and pin-down quantitative definitions of lake quality in the context of trophic states. Therefore, my impression of real-world data met my expectations, for the most part.