Assignment 5: Water Quality in Lakes

Walker Grimshaw

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.

```
# keep warnings from appearing in knitted pdf
knitr::opts_chunk$set(warning = FALSE)

getwd()
```

```
## [1] "C:/Users/walke/OneDrive/Documents/Duke/Courses/Fall_2019/Hydrologic_Data_Analysis/Assignments"
```

```
packages <- c("tidyverse",
              "lubridate",
              "LAGOSNE")
invisible(lapply(packages, library, character.only = TRUE))

theme_set(theme_bw())

## Load LAGOSdata and trophicstate csv
LAGOSdata <- lagosne_load()
```

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

```
LAGOStrophic <- read.csv(file = "../Data/Processed/LAGOStrophic.csv")
```

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).

```
LAGOStrophic <- LAGOStrophic %>%
  mutate(trophic.class.secchi =
```

```
          if_else(TSI.secchi < 40, "Oligotrophic",
                  if_else(TSI.secchi < 50, "Mesotrophic",
                          if_else(TSI.secchi < 70, "Eutrophic",
                                  "Hypereutrophic"))),
        trophic.class.tp =
          if_else(TSI.tp < 40, "Oligotrophic",
                  if_else(TSI.tp < 50, "Mesotrophic",
                          if_else(TSI.tp < 70, "Eutrophic",
                                  "Hypereutrophic"))))

## make new trophic class columns into factors instead of strings
LAGOStrophic$trophic.class.secchi <- as.factor(LAGOStrophic$trophic.class.secchi)
LAGOStrophic$trophic.class.tp <- as.factor(LAGOStrophic$trophic.class.tp)
```

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.

```
## create count tables
CHL.trophic.count <- count(LAGOStrophic, trophic.class)
secchi.trophic.count <- count(LAGOStrophic, trophic.class.secchi)
tp.trophic.count <- count(LAGOStrophic, trophic.class.tp)

## combine count tables
trophic.count <- cbind(CHL.trophic.count,
                       secchi.trophic.count$n,
                       tp.trophic.count$n) %>%
  rename(chlorophyll = n,
         secchi = "secchi.trophic.count$n",
         tp = 'tp.trophic.count$n')
trophic.count
```

```
##    trophic.class chlorophyll secchi    tp
## 1      Eutrophic       41861  28659 24839
## 2 Hypereutrophic       14379   5099  7228
## 3    Mesotrophic       15413  25083 23023
## 4   Oligotrophic        3298  16110 19861
```

7. What proportion of total observations are considered eutrophic or hypereutrophic according to the three different metrics (trophic.class, trophic.class.secchi, trophic.class.tp)?

```
trophic.proportion <- trophic.count %>%
  mutate(chlorophyll = round(chlorophyll/sum(chlorophyll),3),
         secchi = round(secchi/sum(secchi),3),
         tp = round(tp/sum(tp),3))
trophic.proportion
```

```
##    trophic.class chlorophyll secchi    tp
## 1      Eutrophic       0.559  0.382 0.331
## 2 Hypereutrophic       0.192  0.068 0.096
## 3    Mesotrophic       0.206  0.335 0.307
## 4   Oligotrophic       0.044  0.215 0.265
```

Which of these metrics is most conservative in its designation of eutrophic conditions? Why might this be?

> Total phosphorus is the most conservative metric for eutrophic designation. One reason for this may be that when a large amount of phosphorus enters a lake, some of it is used for biomass production in the lake, reducing the concentration of measured phosphorus, though the biomass

production indicates a higher trophic state.

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.

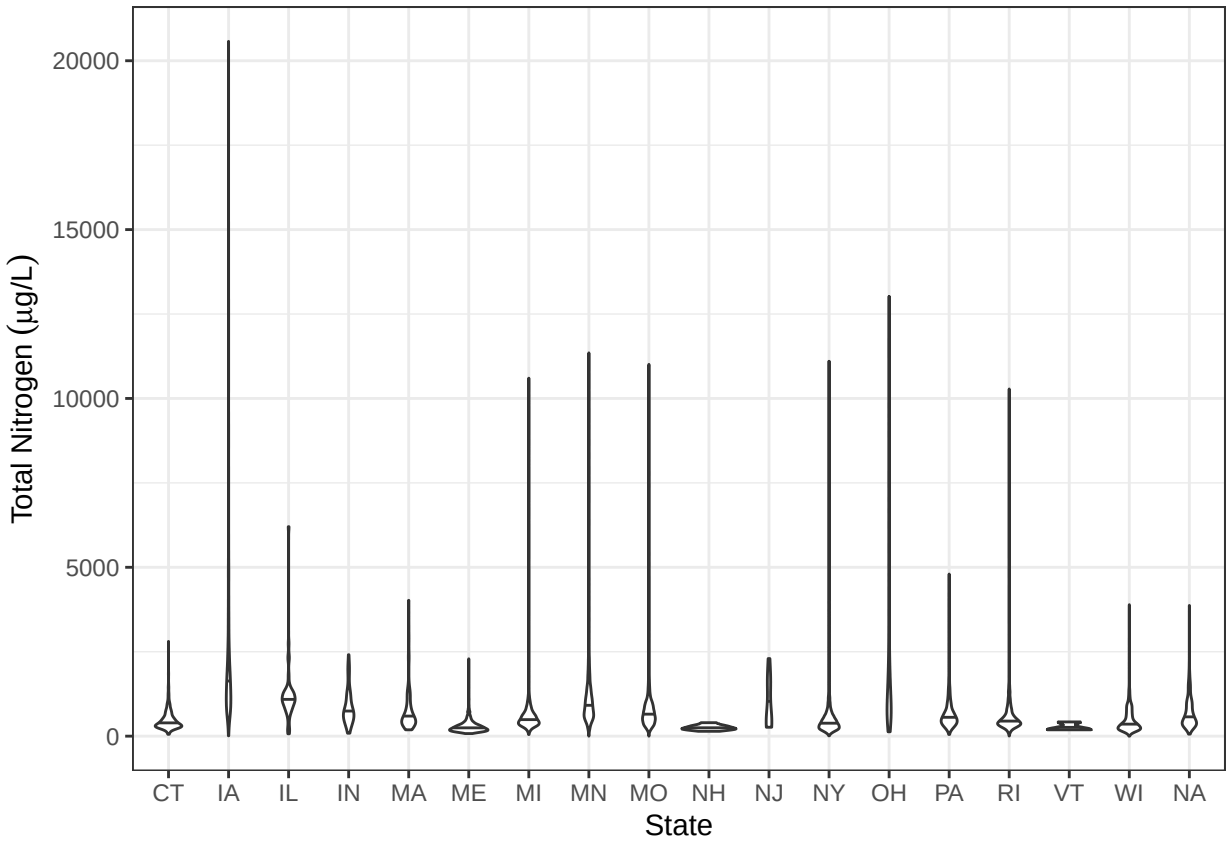
Nutrient Concentrations

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.

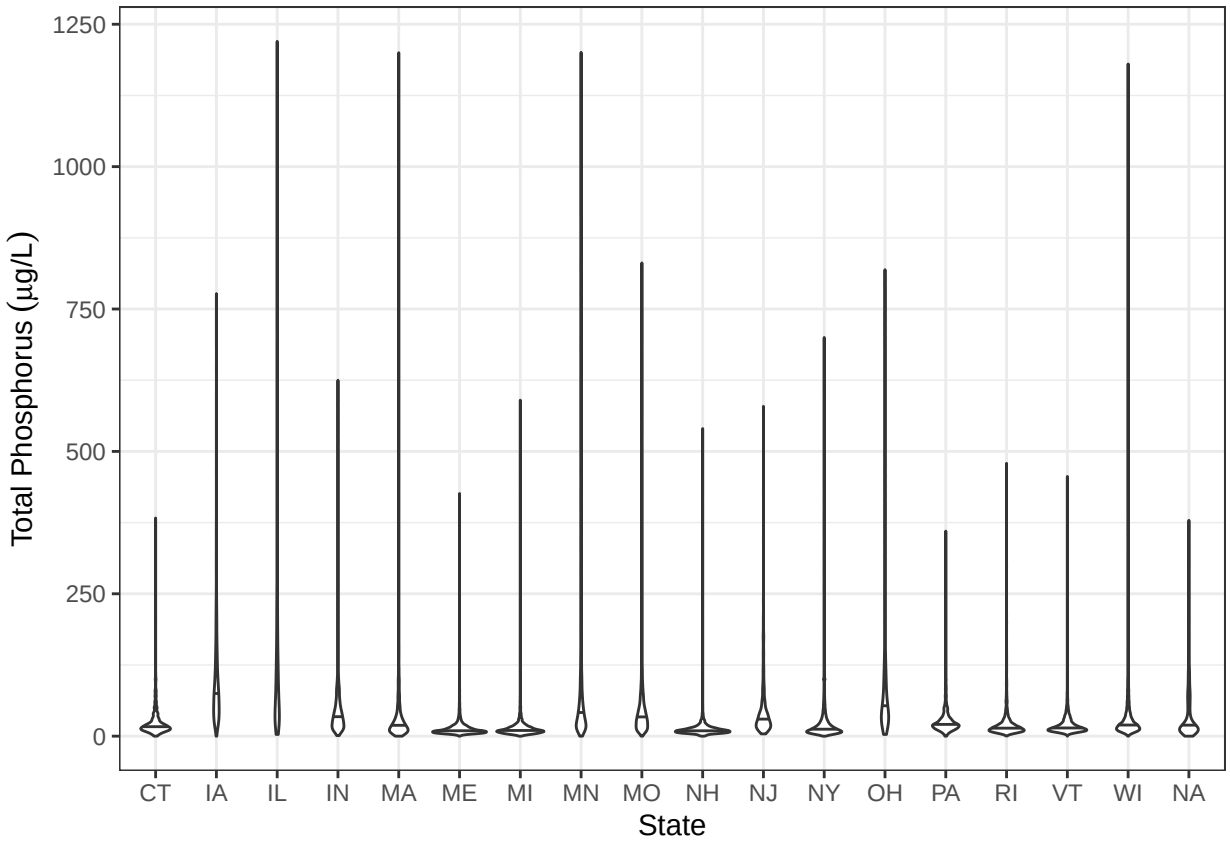
```
## load nutrient, locus and state data
LAGOSnutrients <- LAGOSdata$epi_nutr %>%
  select(lagoslakeid, sampledate, tn, tp)
LAGOSlocus <- LAGOSdata$locus
LAGOSstate <- LAGOSdata$state

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

## create N and P data frame
LAGOSNandP <-
  left_join(LAGOSnutrients, LAGOSlocations, by = "lagoslakeid") %>%
  select(lagoslakeid, sampledate, tn, tp, state, state_name) %>%
  ## remove sample points missing both n and p data
  filter(is.na(tn) == FALSE | is.na(tp) == FALSE) %>%
  mutate(sampleyear = year(sampledate),
         samplemonth = month(sampledate))
```

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

```
ggplot(data = LAGOSNandP, aes(x = state, y = tn)) +
  geom_violin(draw_quantiles = 0.5) +
  labs(x = "State", y = expression("Total Nitrogen" ~ (mu*"g/L")))
```




```
ggplot(data = LAGOSNandP, aes(x = state, y = tp)) +
  geom_violin(draw_quantiles = 0.5) +
  labs(x = "State", y = expression("Total Phosphorus" ~ (mu*"g/L")))
```




```
LAGOS.summary <- LAGOSNandP %>%
  group_by(state) %>%
  summarize(median.tn = median(tn, na.rm = T), median.tp = median(tp, na.rm = T),
            range.tn = max(tn, na.rm = T) - min(tn, na.rm = T),
            range.tp = max(tp, na.rm = T) - min(tp, na.rm = T))
```

Which states have the highest and lowest median concentrations?

TN: The state with the highest median concentration is Iowa and the lowest is Vermont.

TP: The state with the highest median concentration is Illinois and the lowest is a tie between New Hampshire and Maine.

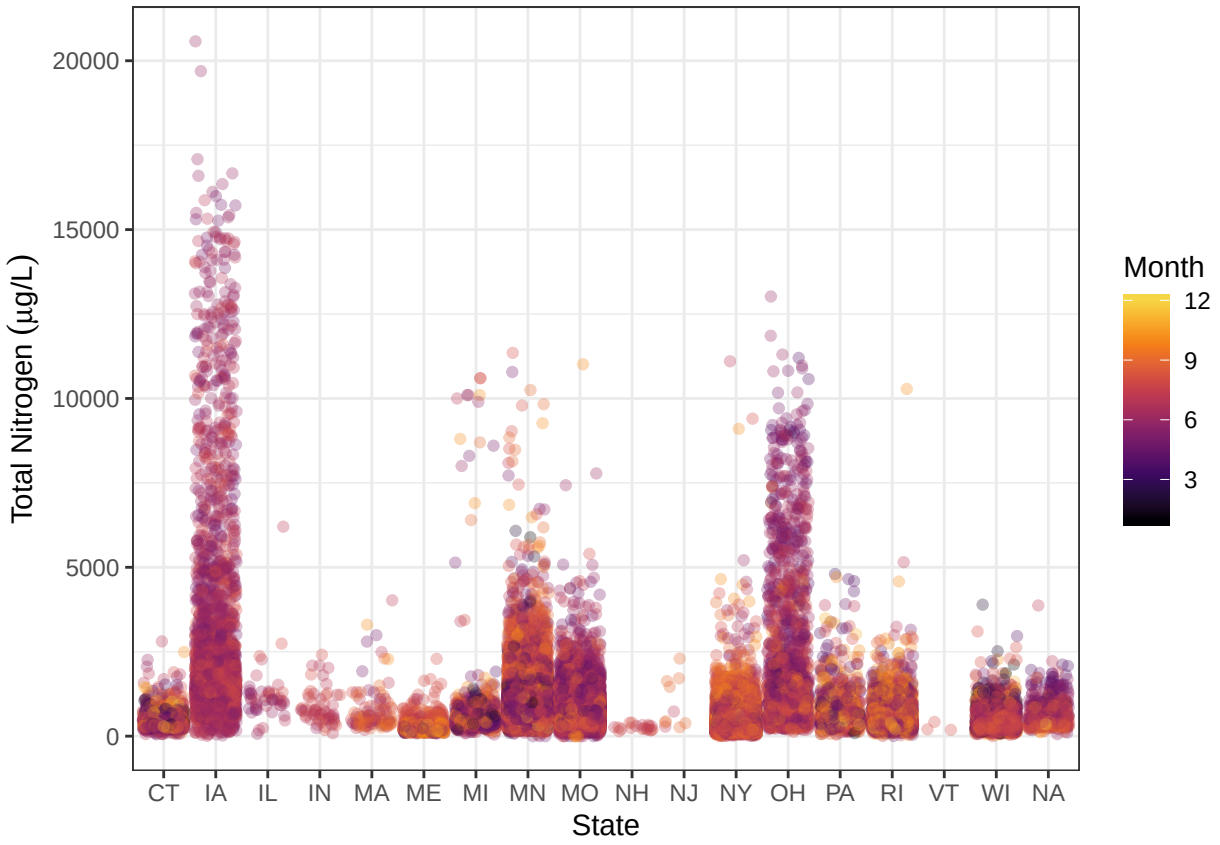
Which states have the highest and lowest concentration ranges?

TN: The highest range is in Iowa and the lowest range is in Vermont.

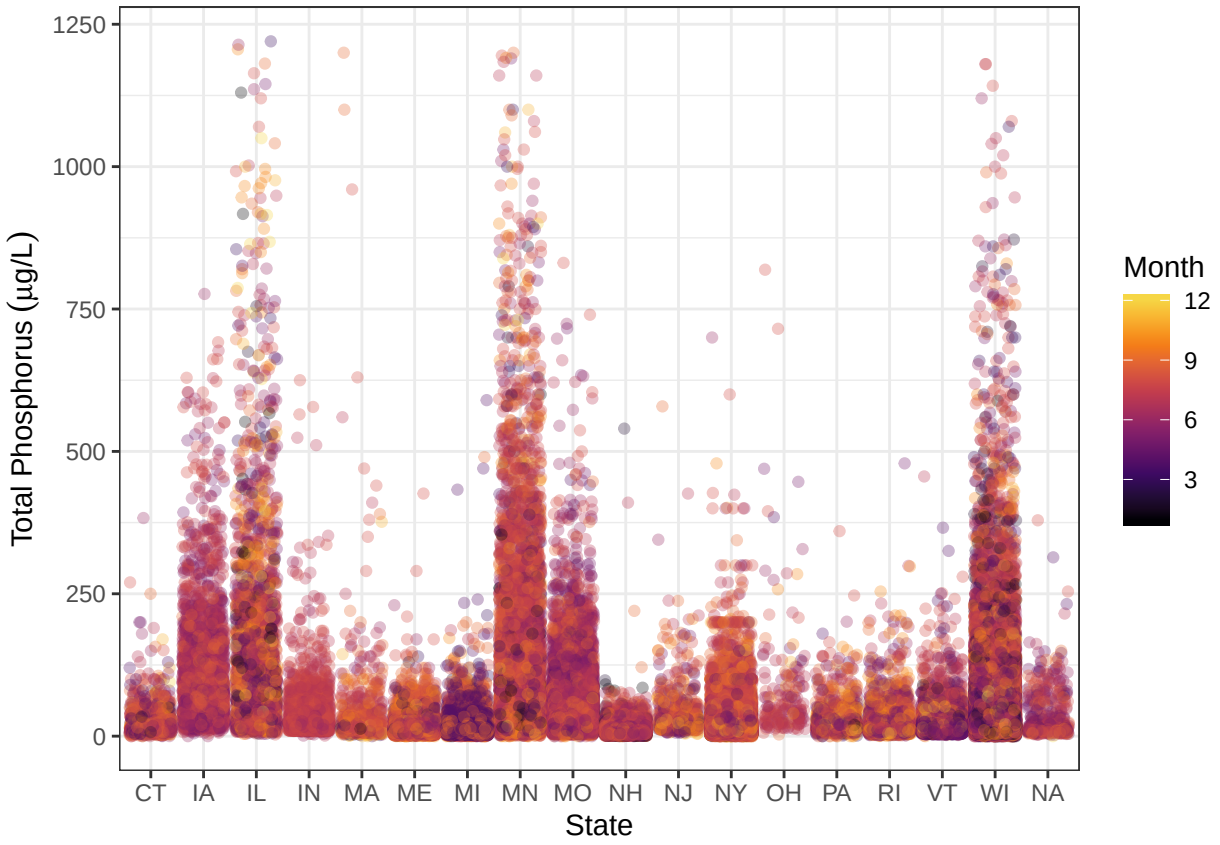
TP: The highest range is in Illinois and the lowest range is in Pennsylvania.

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.

```
ggplot(data = LAGOSNandP, aes(x = state, y = tn, color = samplemonth)) +
  geom_jitter(alpha = 0.3) +
  labs(x = "State", y = expression("Total Nitrogen" ~ (mu*"g/L")),
       color = "Month") +
  scale_color_viridis_c(option = "inferno", end = 0.9)
```




```
ggplot(data = LAGOSNandP, aes(x = state, y = tp, color = samplemonth)) +
  geom_jitter(alpha = 0.3) +
  labs(x = "State", y = expression("Total Phosphorus" ~ (mu*"g/L")),
       color = "Month") +
  scale_color_viridis_c(option = "inferno", end = 0.9)
```




```
state.tn <- LAGOSNandP %>%
  select(state, tn) %>%
  filter(is.na(tn) == FALSE) %>%
  count(state)

state.tp <- LAGOSNandP %>%
  select(state, tp) %>%
  filter(is.na(tp) == FALSE) %>%
  count(state)
```

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

> TN: Missouri, New York, Rhode Island, and Iowa have the most sample points. Iowa does seem to be an outlier in its range though, as New York and Missouri have far more observations. Iowa's range and high median may be more indicative of the large agricultural industry there.
>
> TP: Wisconsin, New York, and the M-states besides Massachusetts have the most sample points. These do correspond with the states that have some of the largest ranges. This makes sense, as a greater sample size would likely lead to a greater chance of observing extreme values.

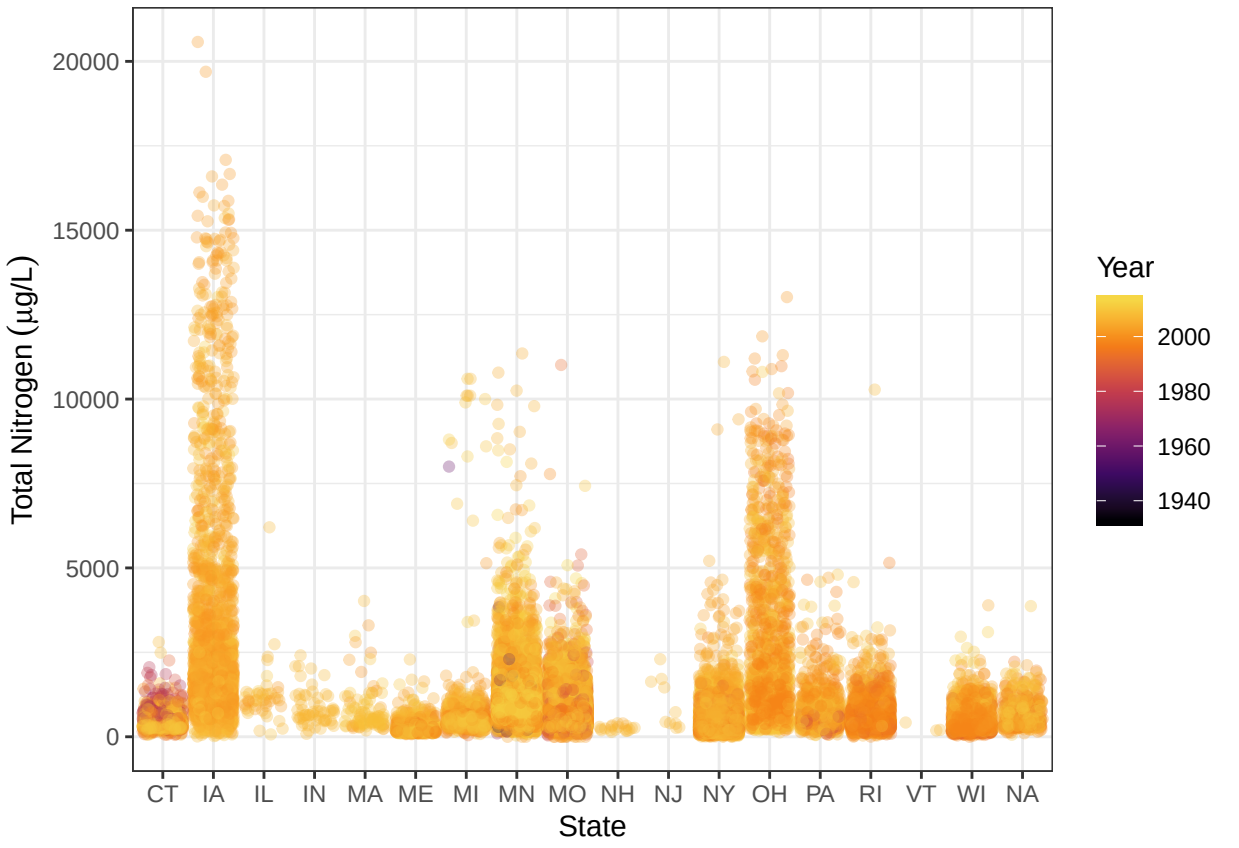
Which months are sampled most extensively? Does this differ among states?

> TN: The months sampled most extensively do differ by state. Iowa and Ohio appear to sample most frequently in May through July, while Minnesota and New York sample in the late Summer and early Fall.
>
> TP: As with Nitrogen, the months that are sampled most extensively do differ by state. States like Iowa, Indiana, and Missouri sample most in June and July, but Minnesota and New York sample slightly lighter in the season, largely in August and September. Mississippi appears to

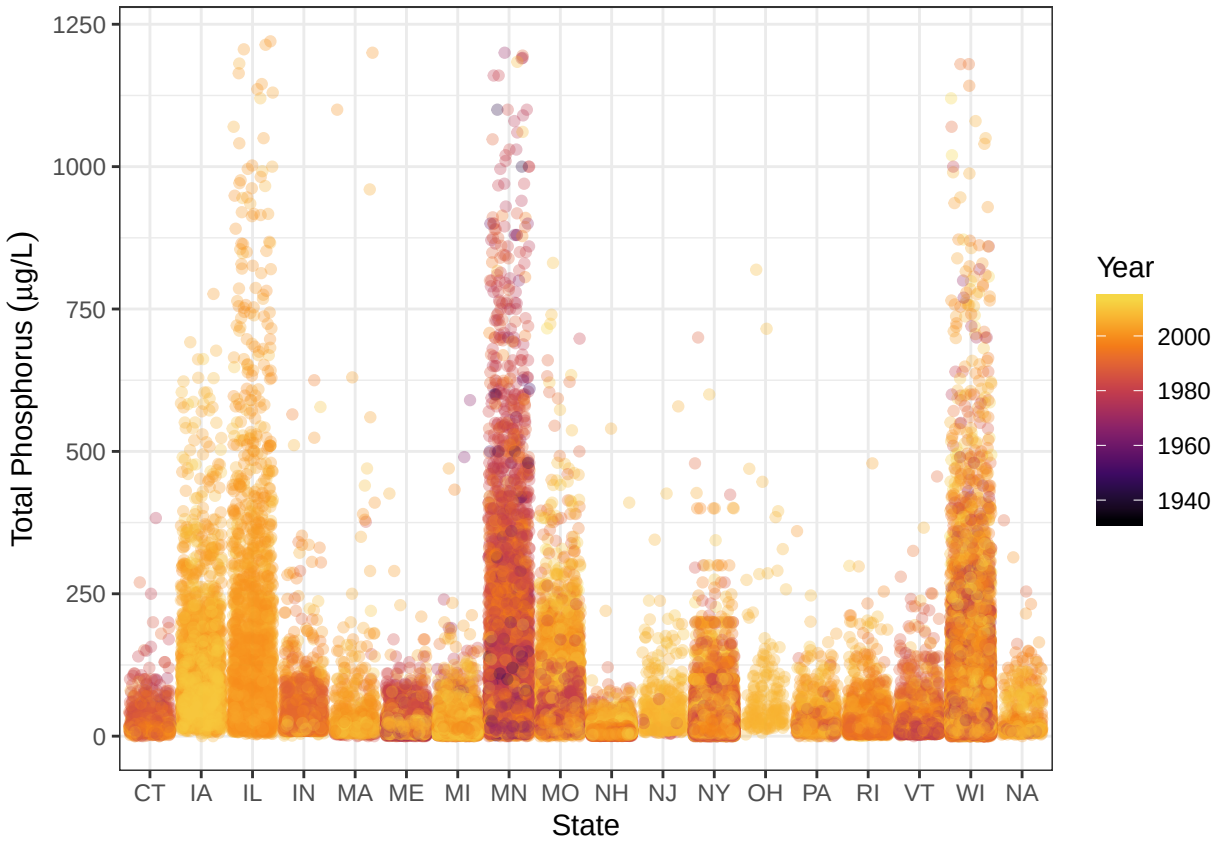
have sampled extensively in winter and early spring months.

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.

```
ggplot(data = LAGOSNandP, aes(x = state, y = tn, color = sampleyear)) +
  geom_jitter(alpha = 0.3) +
  labs(x = "State", y = expression("Total Nitrogen" ~ (mu*"g/L")),
       color = "Year") +
  scale_color_viridis_c(option = "inferno", end = 0.9)
```




```
ggplot(data = LAGOSNandP, aes(x = state, y = tp, color = sampleyear)) +
  geom_jitter(alpha = 0.3) +
  labs(x = "State", y = expression("Total Phosphorus" ~ (mu*"g/L")),
       color = "Year") +
  scale_color_viridis_c(option = "inferno", end = 0.9)
```



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

TN: Almost all states sample most extensively after the late 1990s. This seems to vary only slightly by state, with states such as Ohio, Rhode Island, and Wisconsin sampling a few years earlier than most other states.

TP: Phosphorus sampling differs substantially from nitrogen sampling. Minnesota has far more samples from before 1990 than any other state, with only Connecticut following a similar pattern of early sampling. Most other states have sample mostly after 2000, but states still show more variation than for nitrogen sampling.

Reflection

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

    Trophic states may be defined by different metrics, and the delineation between trophic states is not always precise. However, the trophic state of a lake is related to nutrient loading, water clarity, and chlorophyll concentration, all used as proxies for total lake biomass. Additionally, states vary greatly in their monitoring of lakes for nutrients of interest.

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

    When different metrics are used, a lake may be deemed to be in a different trophic class, as indicated by the tables in questions (6) and (7). The four jitter plots show the difference in lake monitoring frequency by state and the differences in how long they have been monitoring their lakes for nutrients that may determine trophic state.

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

The data analysis gave a better picture of the inconsistency of trophic state assignment. Also, the visualizations by state make me want to further investigate the funding mechanisms and policies that control lake nutrient monitoring across different states.

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

    There were substantial differences between the three different trophic state indices. I am curious if an aggregate index of the three would provide a more consistent index.