Assignment 5: Water Quality in Lakes

Haoyu Zhang

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.

```
knitr::opts_chunk$set(echo = TRUE, warning = F)

getwd()
```

```
## [1] "D:/William/Duke/Study/EOS 722/Hydrologic_Data_Analysis/Assignments/Assignment 5"
```

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

```
## -- Attaching packages ---------------------------------------------------- tidyverse 1.2.1 --
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     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()
##
## Attaching package: 'lubridate'
## The following object is masked from 'package:base':
##
##     date
```

```
theme_set(theme_classic())
options(scipen = 100)
LAGOSdat <- load(file = "../../Data/Raw/LAGOSdata.rda")
# NOTE need to change the path relative to the rmd location when knitting rmarkdown; it does not recogn
```

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).

```
# Retrieve useful dataframes from metadata
LAGOSlocus <- LAGOSdata$locus
LAGOSstate <- LAGOSdata$state
LAGOSnutrient <- LAGOSdata$epi_nutr

# Join locus with state, giving a location dataframe
LAGOSlocations <- left_join(LAGOSlocus, LAGOSstate, by = "state_zoneid")

# Join nutrient with location
LAGOStrophic <-
  left_join(LAGOSnutrient, LAGOSlocations, by = "lagoslakeid") %>%
  select(lagoslakeid, sampledate, chla, tp, secchi,
         gnis_name, lake_area_ha, state, state_name) %>%
  mutate(sampleyear = year(sampledate),
         samplemonth = month(sampledate),
         season = as.factor(quarter(sampledate, fiscal_start = 12))) %>%
  # Dec, Jan, Feb - winter (1), Mar, April, May - Spring (2), and so on
  drop_na(chla:secchi)

levels(LAGOStrophic$season) <- c("Winter", "Spring", "Summer", "Fall")

# Determine trophic status by chla, secchi, and tp
LAGOStrophic <-
  mutate(LAGOStrophic,
         TSI.chl = round(10*(6 - (2.04 - 0.68*log(chla)/log(2)))),
         TSI.secchi = round(10*(6 - (log(secchi)/log(2)))),
         TSI.tp = round(10*(6 - (log(48/tp)/log(2)))),
         # Trophic status by chla
         trophic.class.chl =
            if_else(TSI.chl < 40, "Oligotrophic",
                   if_else(TSI.chl < 50, "Mesotrophic",
                          if_else(TSI.chl < 70, "Eutrophic", "Hypereutrophic"))),
         # Trophic status by secchi
         trophic.class.secchi =
            if_else(TSI.secchi < 40, "Oligotrophic",
                   if_else(TSI.secchi < 50, "Mesotrophic",
                          if_else(TSI.secchi < 70, "Eutrophic", "Hypereutrophic"))),
         # Trophic status by tp
         trophic.class.tp =
            if_else(TSI.tp < 40, "Oligotrophic",
                   if_else(TSI.tp < 50, "Mesotrophic",
                          if_else(TSI.tp < 70, "Eutrophic", "Hypereutrophic")))
         )

# Convert character vectors to factors
LAGOStrophic$trophic.class.chl <-
  factor(LAGOStrophic$trophic.class.chl,
         levels = c("Oligotrophic", "Mesotrophic", "Eutrophic", "Hypereutrophic"))
LAGOStrophic$trophic.class.secchi <-
```

```
  factor(LAGOStrophic$trophic.class.secchi,
         levels = c("Oligotrophic", "Mesotrophic", "Eutrophic", "Hypereutrophic"))
LAGOStrophic$trophic.class.tp <-
  factor(LAGOStrophic$trophic.class.tp,
         levels = c("Oligotrophic", "Mesotrophic", "Eutrophic", "Hypereutrophic"))
# LAGOStrophic$season <-
#   factor(LAGOStrophic$season,
#          levels = c("Spring", "Summer", "Fall", "Winter"))
```

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.chl <- count(LAGOStrophic, LAGOStrophic$trophic.class.chl)
count.secchi <- count(LAGOStrophic, LAGOStrophic$trophic.class.secchi)
count.tp <- count(LAGOStrophic, LAGOStrophic$trophic.class.tp)
```

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

```
sum(count.chl$n[3:4]) / sum(count.chl$n) # Proportion according to chl
```

```
## [1] 0.7503569
```

```
sum(count.secchi$n[3:4]) / sum(count.secchi$n) # Proportion according to secchi
```

```
## [1] 0.4504009
```

```
sum(count.tp$n[3:4]) / sum(count.tp$n) # Proportion according to tp
```

```
## [1] 0.4278395
```

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

> Total phosphorus is the most conservative, as it has the lowest proportion among the three (0.43). The tendency to underestimate the trophic status suggests that phosphorus is low even in eutrophic/hypereutrophic lakes. One plausible reason could be that available phosphorus can be rapidly removed from the nutreitn pool by algae/other organisms, leaving low proportion remained in soluation.

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.

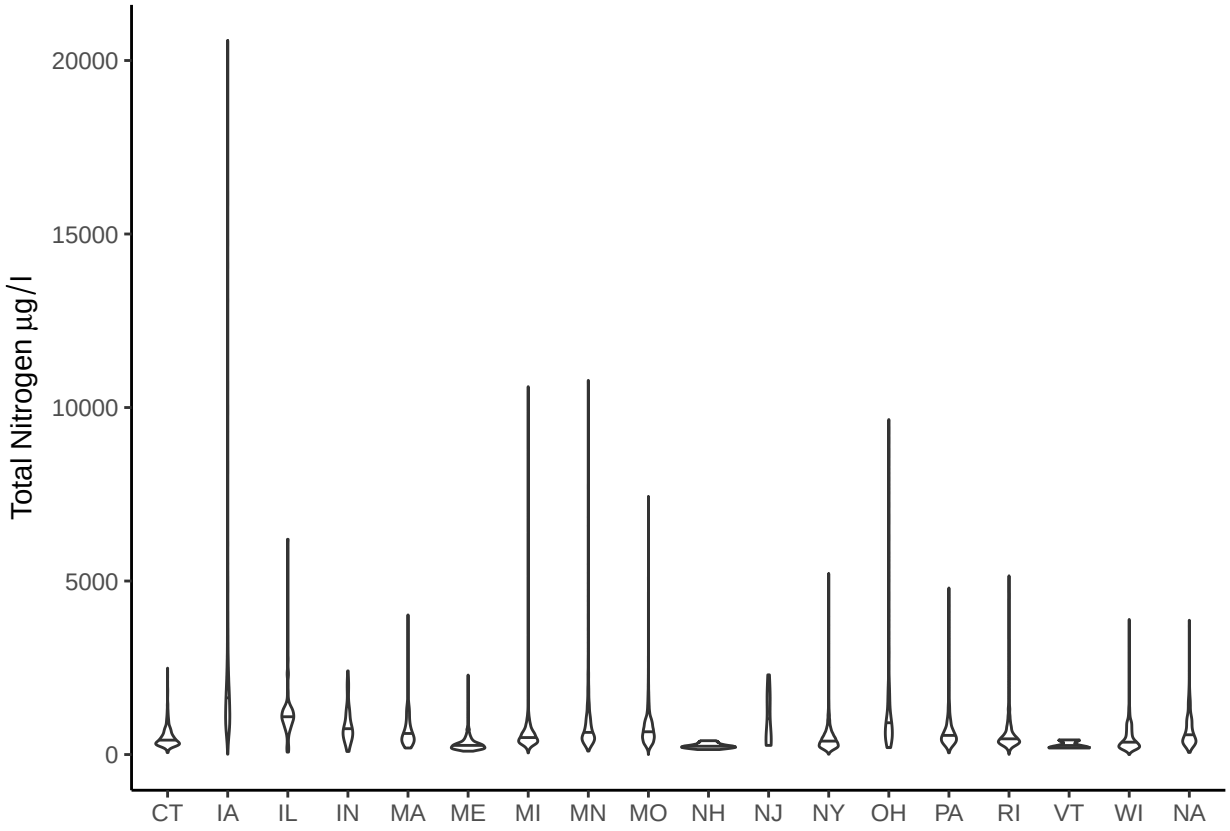
Nutrient Concentrations

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.

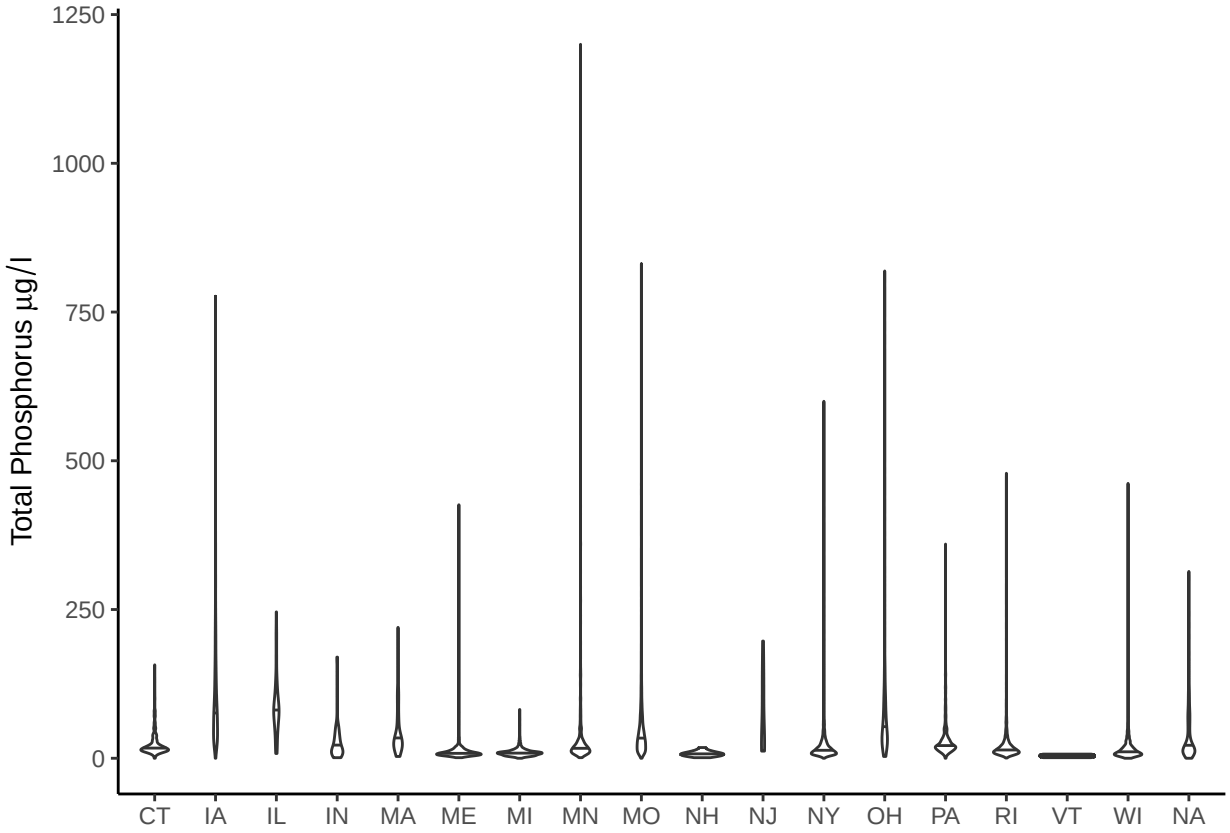
```
LAGOSNandP <- left_join(LAGOSnutrient, LAGOSlocations, by = "lagoslakeid") %>%
  select(lagoslakeid, sampledate, tn, tp, state, state_name) %>%
  mutate(sampleyear = year(sampledate),
         samplemonth = month(sampledate),) %>%
  drop_na(c(tn, tp))
```

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

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




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




```
# Log transformed y-scale; NOTE relative range changes
#ggplot(LAGOSNandP, aes(y = tn, x = state))+
#  geom_violin(draw_quantiles = 0.5)+
#  scale_y_log10(name = expression("Total Nitrogen "*mu*g/l))+
#  annotation_logticks(base = 10, scaled = T, side = "l")+
#  labs(x = element_blank())

#ggplot(LAGOSNandP, aes(y = tp, x = state))+
#  geom_violin(draw_quantiles = 0.5)+
#  scale_y_log10(name =  expression("Total Phosphorus "*mu*g/l))+
#  annotation_logticks(base = 10, scaled = T, side = "l")+
#  labs(x = element_blank())
```

Which states have the highest and lowest median concentrations?

TN: highest - IA; lowest - VT

TP: highest - IL; lowest - VT

Which states have the highest and lowest concentration ranges?

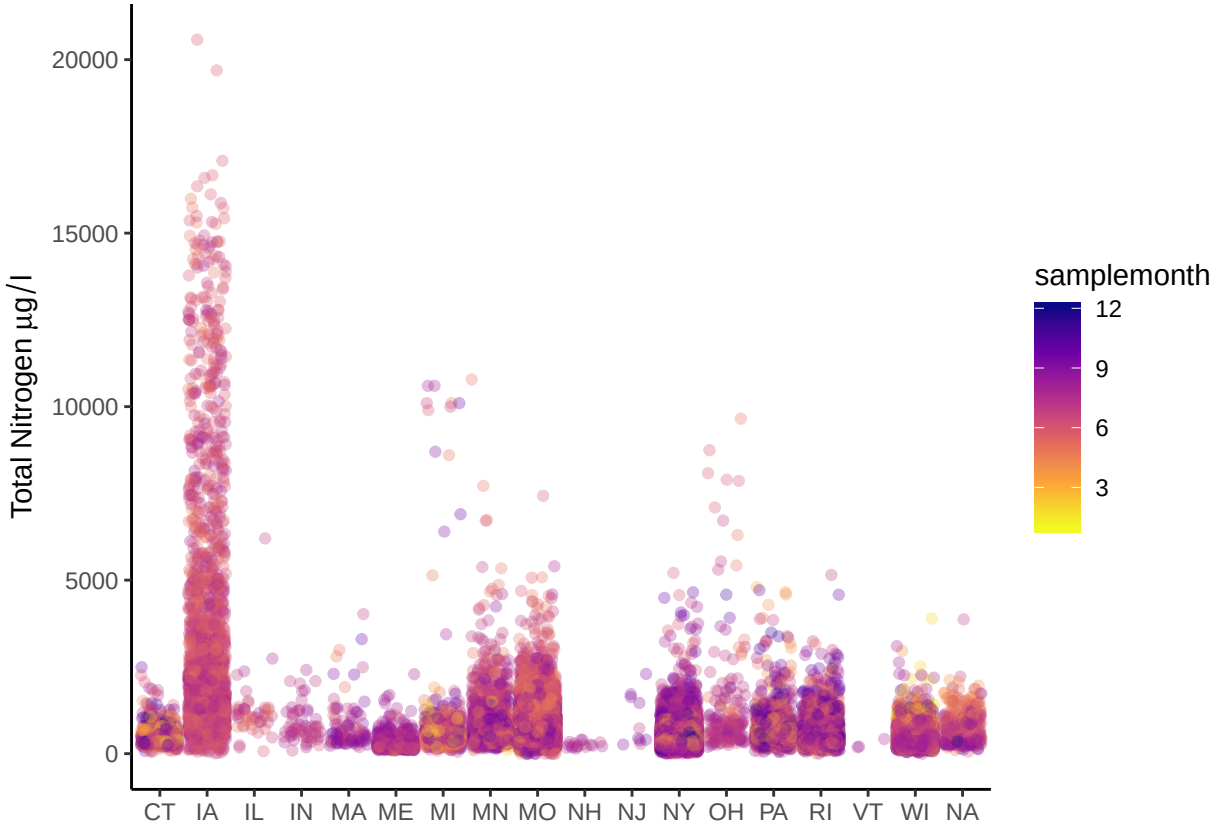
TN: highest - IA; lowest - VT

TP: highest - MN; lowest - VT

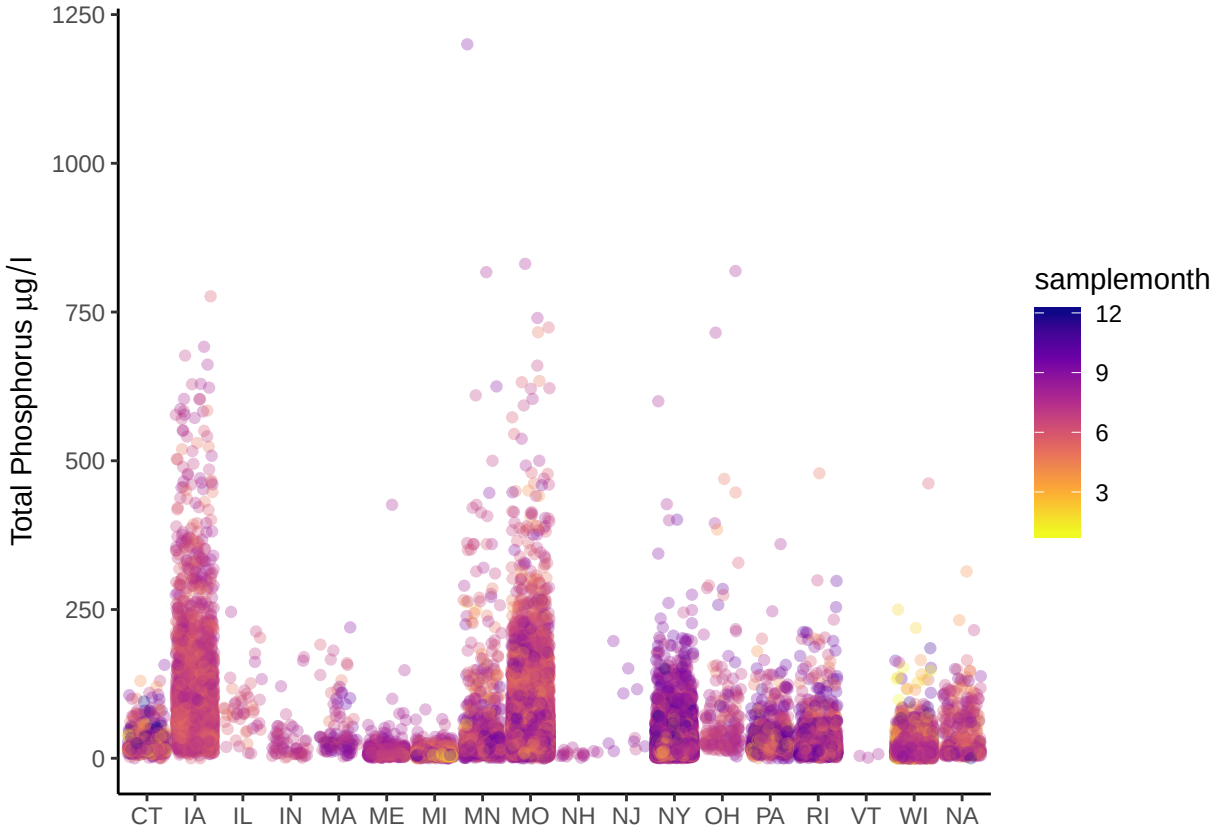
10. Create two jitter plots comparing TN and TP concentrations across states, with samplemonth as the color. Choose a color palette other than the ggplot default.

```
ggplot(LAGOSNandP, aes(x = state, y = tn, col = samplemonth))+
  geom_jitter(alpha = 0.3)+
```

```
  scale_color_viridis_c(option = "plasma", direction = -1)+
  labs(x = element_blank(), y = expression("Total Nitrogen "*mu*g/l))
```




```
ggplot(LAGOSNandP, aes(x = state, y = tp, col = samplemonth))+
  geom_jitter(alpha = 0.3)+
  scale_color_viridis_c(option = "plasma", direction = -1)+
  labs(x = element_blank(), y = expression("Total Phosphorus "*mu*g/l))
```



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

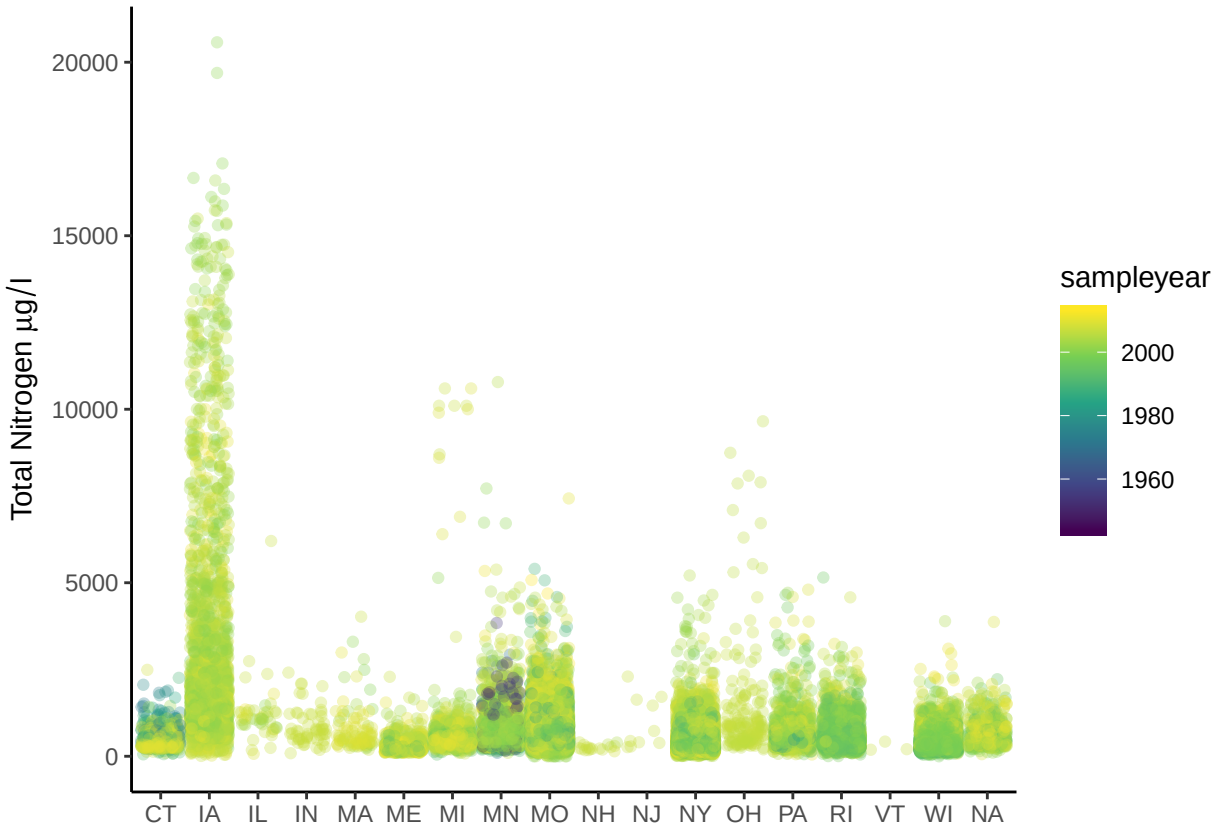
> TN: IA has the most samples, and is also the state with the largest sample range. As more sampling effort is devoted in a region, the variation among lakes is more likely to be detected, resulting in a larger range.
>
> TP: Same trend can be observed for TP: MN has the most samples and the largest range.

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

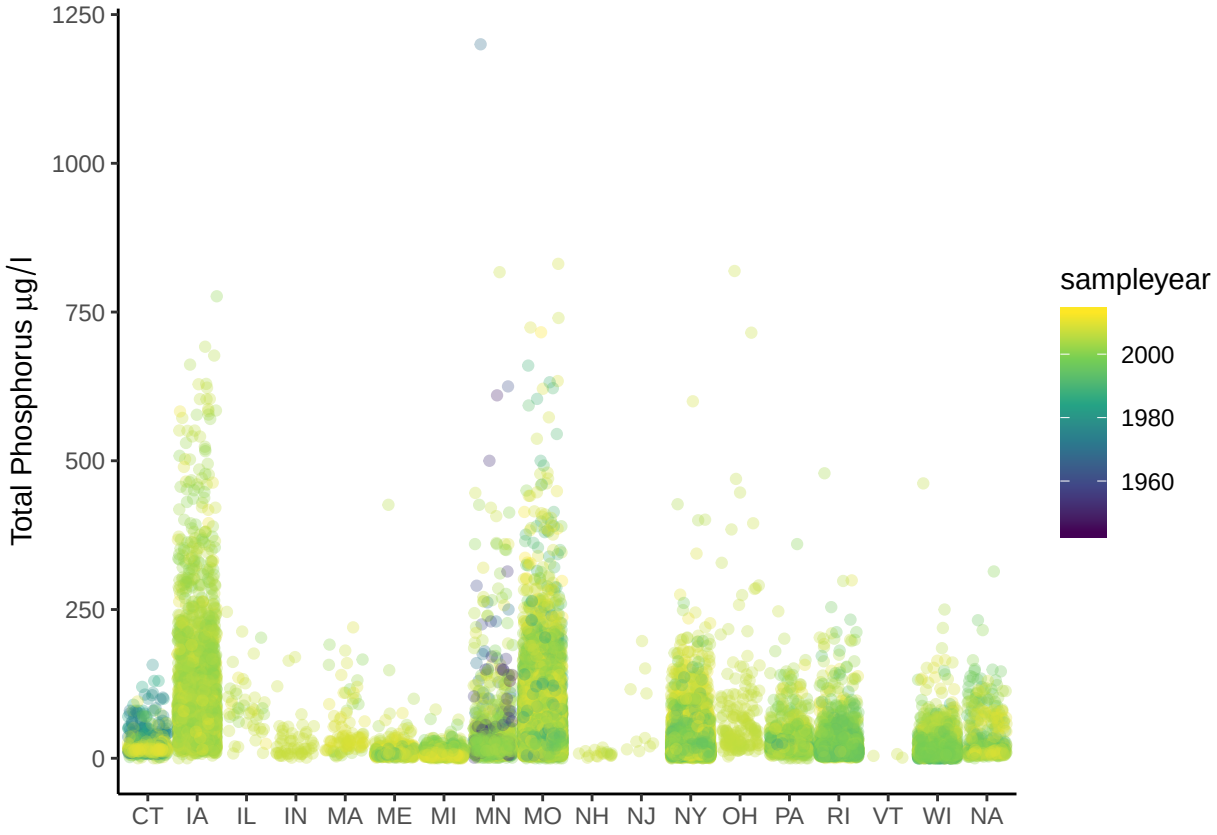
> TN: Sep to Dec are most extensively sampled. There is difference to some extent but no considerable among states. Some states, such as IA, MO, WI, NA, have slightly more samples during the first half of the year.
>
> TP: Similar to TP, the second half of the year is the time when most measurements were taken, and the distribution of sample date differs moderately among states.

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(LAGOSNandP, aes(x = state, y = tn, col = sampleyear))+
  geom_jitter(alpha = 0.3)+
  scale_color_viridis_c(option = "viridis", direction = 1)+
  labs(x = element_blank(), y = expression("Total Nitrogen "*mu*g/l))
```




```
ggplot(LAGOSNandP, aes(x = state, y = tp, col = sampleyear))+
  geom_jitter(alpha = 0.3)+
  scale_color_viridis_c(option = "viridis", direction = 1)+
  labs(x = element_blank(), y = expression("Total Phosphorus "*mu*g/l))
```



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

TN: Samples were taken predominantly after 1980, but notably CT and MN have much more observations before 1980 than the others. A few states, e.g. NH, NJ VT, only have observations after about 2000.

TP: Very similar pattern exists for TP. Most states only have observations after 1980, but CT and MN have pre-1980 measurements. Some states only have data in the last few years.

Reflection

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

    First, there could arise discrepancy in determining the trophic status of a water body among different metrics (e.g. chlorophyll, secchi depth, and total phosphorus). Among the three mentioned here, total phosphorus is the most conservative metric in terms of designating eutrophic/hypereutrophic. Secondly, the variation among a sample set collected from a region can heavily depend on the sampling effort that has been put into assessing water bodies. Regions with more samples are likely to have larger variations. This may call attention to sample size while comparing water quality variables across regions.

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

    The first point was drawn based on the calculating proportions of (hyper)eutrophic lakes based on different metrics. The second conclusion is based on the comparison between violin plots and jitter plots.

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

Hands-on analysis definitely helps understand different veridicts on trophic status given by different metrics. The analyses on the potential bias on variance caused by sample size could serve as an empirical evidence for the caveat that comparably equal sample size is important and needs to be met if we want to compared across regions (and possibly for statistical purposes).

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

    The tendency of phosphorus to underestimate the proportion of eutrophic lakes is somewhat unexpected to me, but the strong predictive ability of chlorophyll concentration on trophic status is in accordance with my expectation, since it can be directly linked to algal bloom.