Assignment 5: Water Quality in Lakes

Caroline Watson

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.

```
#checking working directory
getwd()
```

```
## [1] "/Users/carolinewatson/Documents/Fall 2019/Hydrologic_Data_Analysis/Assignments"
```

```
#loading packages
suppressMessages(library(tidyverse))
library(lubridate)
```

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

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

```
library(LAGOSNE)
library(viridis)
```

```
## Loading required package: viridisLite
```

```
#setting ggplot theme
theme_set(theme_classic())
options(scipen = 100)

#loading LAGOS dataset
LAGOSdata <- lagosne_load()
```

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

```
names(LAGOSdata)
```

```
##  [1] "county"               "county.chag"          "county.conn"
##  [4] "county.lulc"          "edu"                  "edu.chag"
##  [7] "edu.conn"             "edu.lulc"             "hu4"
## [10] "hu4.chag"             "hu4.conn"             "hu4.lulc"
## [13] "hu8"                  "hu8.chag"             "hu8.conn"
## [16] "hu8.lulc"             "hu12"                 "hu12.chag"
## [19] "hu12.conn"            "hu12.lulc"            "iws"
## [22] "iws.conn"             "iws.lulc"             "state"
## [25] "state.chag"           "state.conn"           "state.lulc"
## [28] "buffer100m"           "buffer100m.lulc"      "buffer500m"
## [31] "buffer500m.conn"      "buffer500m.lulc"      "lakes.geo"
## [34] "epi_nutr"             "lakes_limno"          "lagos_source_program"
## [37] "locus"
```

```
#loading trohic state index csv file
LAGOS.trophic <- read.csv("../Data/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).

```
#creating trophic.class.secchi column in datafram
LAGOStrophic <-
  mutate(LAGOS.trophic,
         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.class =
            ifelse(TSI.chl < 40, "Oligotrophic",
                   ifelse(TSI.chl < 50, "Mesotrophic",
                          ifelse(TSI.chl < 70, "Eutrophic", "Hypereutrophic"))),
         trophic.class.secchi =
           ifelse(TSI.secchi < 40, "Oligotrophic",
              ifelse(TSI.secchi < 50, "Mesotrophic",
                     ifelse(TSI.secchi < 70, "Eutrophic", "Hypereutrophic"))),
         trophic.class.tp =
           ifelse(TSI.tp < 40, "Oligotrophic",
                  ifelse(TSI.tp < 50, "Mesotrophic",
                         ifelse(TSI.tp < 70, "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.

```
#number of observations in trophic.class
LAGOStrophic %>% count(trophic.class)
```

```
## # A tibble: 4 x 2
##   trophic.class      n
##   <chr>          <int>
```

```
## 1 Eutrophic      41861
## 2 Hypereutrophic 14379
## 3 Mesotrophic    15413
## 4 Oligotrophic    3298
```

```
#number of observations in trophic.class.secchi
LAGOStrophic %>% count(trophic.class.secchi)
```

```
## # A tibble: 4 x 2
##   trophic.class.secchi     n
##   <chr>                <int>
## 1 Eutrophic            28659
## 2 Hypereutrophic        5099
## 3 Mesotrophic          25083
## 4 Oligotrophic         16110
```

```
#number of observations in trophic.class.tp
LAGOStrophic %>% count(trophic.class.tp)
```

```
## # A tibble: 4 x 2
##   trophic.class.tp     n
##   <chr>            <int>
## 1 Eutrophic        24839
## 2 Hypereutrophic    7228
## 3 Mesotrophic      23023
## 4 Oligotrophic     19861
```

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)?

```
#proportion of trophic.class that is eutrophic
(41861/74951)*100 #55.85% which can be rounded to 56%
```

```
## [1] 55.85116
```

```
#proportion of trophic.class that is hypereutrophic
(14379/74951)*100 #19.18% which is about 19%
```

```
## [1] 19.18453
```

```
#proportion of trophic.class.secchi that is eutrophic
(28659/74951)*100 #38.24% which is about 38%
```

```
## [1] 38.23698
```

```
#proportion of trophic.class.secchi that is hypereutrophic
(5099/74951)*100 #6.8% of trophic.class.secchi is hyperutrophic
```

```
## [1] 6.803111
```

```
#proportion of trophic.class.tp that is eutrophic
(24839/74951)*100 #33% of the trophic.class.tp is eutrophic
```

```
## [1] 33.14032
```

```
#proportion of trophic.class.tp that is hypereutrophic
(7228/74951)*100 #9.6% so about 10% of trophic.class.tp is hypereutrophic
```

```
## [1] 9.643634
```

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

Trophic class total phosphorus is the most conservative with its designation of eutrophic conditions. This is likely because there are other things impacting eutrophication, such as secchi depth and total nitrogen. Total phosphorus can limit growth of algae, especially in summer months.

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.

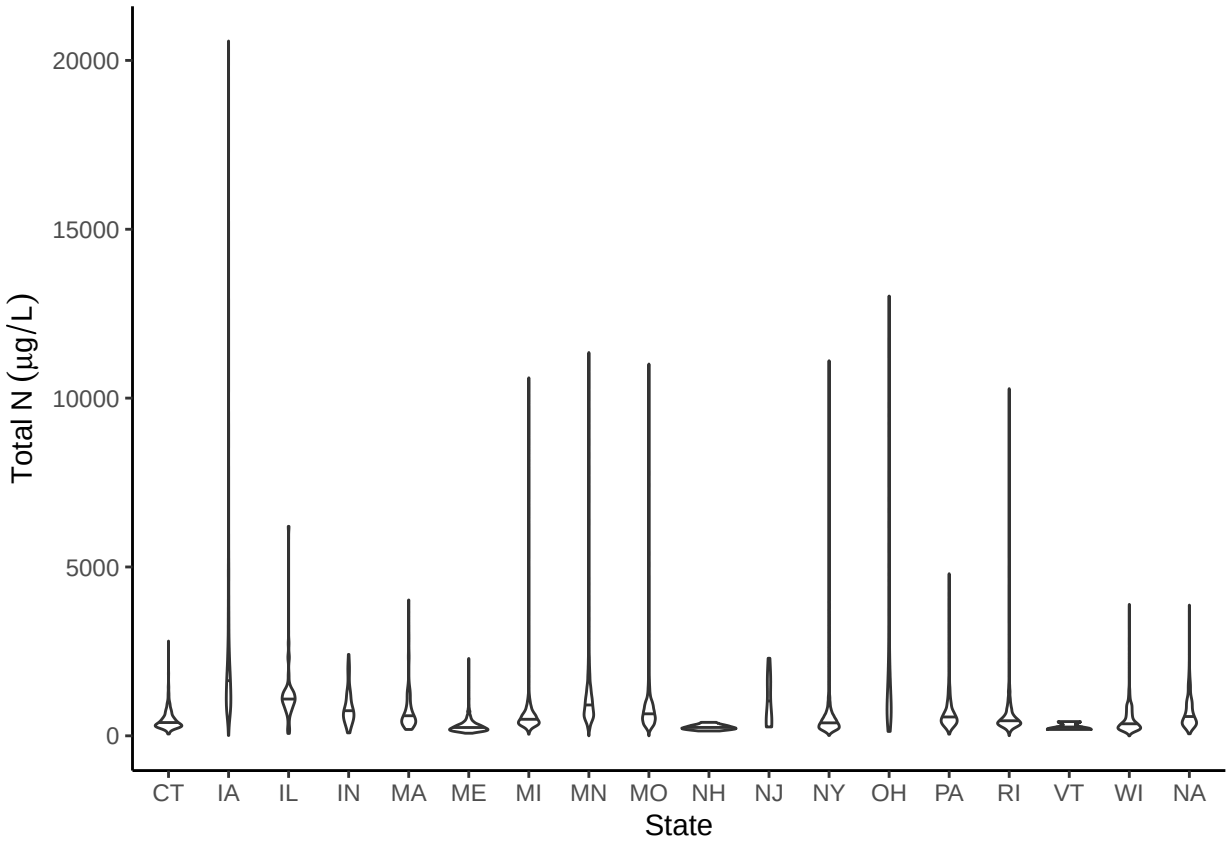
Nutrient Concentrations

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.

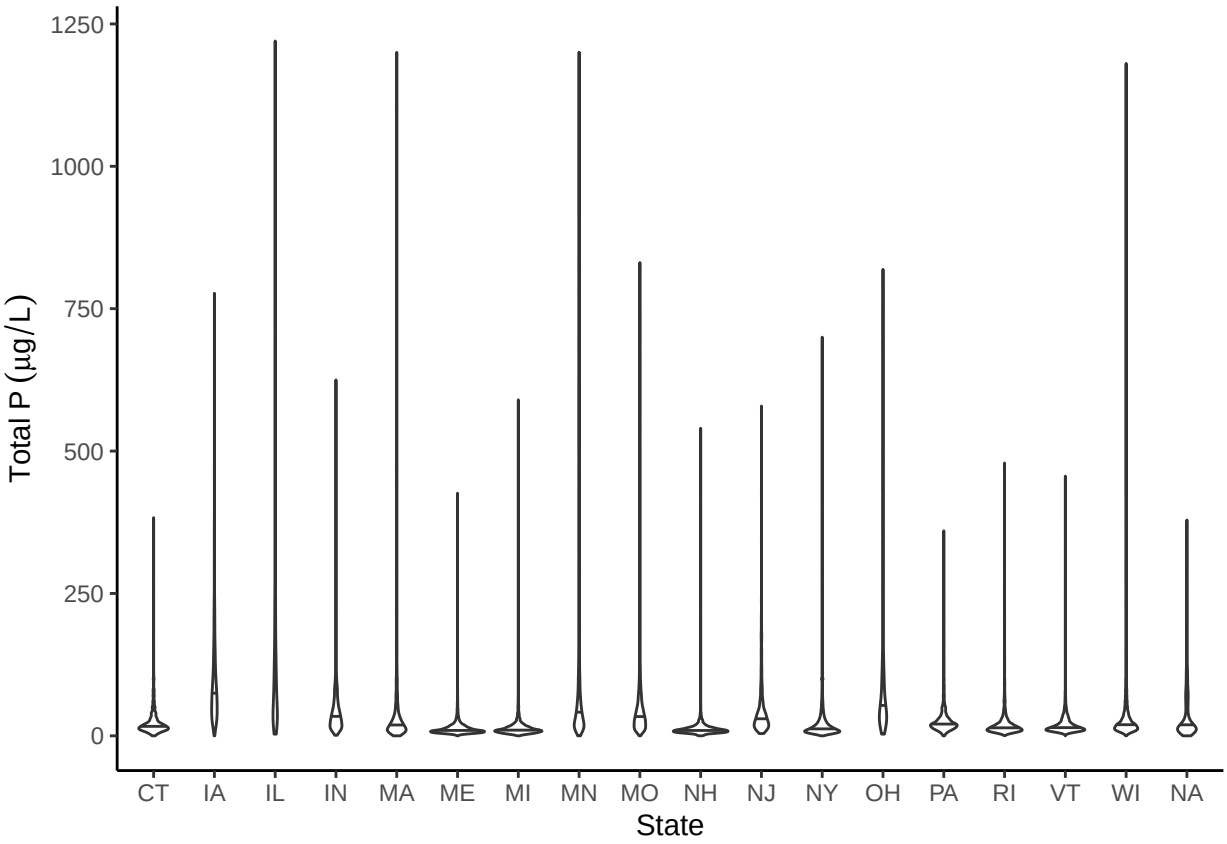
```
#creating LAGOS dataframes
LAGOSlocus <- LAGOSdata$locus
LAGOSstate <- LAGOSdata$state
LAGOSnutrient <- LAGOSdata$epi_nutr

#joining state and nutrient dataframes
LAGOSlocations <- left_join(LAGOSlocus, LAGOSstate, by = "state_zoneid")

#joining LAGOS locations and nutrient dataframs
LAGOSNandP <- left_join(LAGOSlocations, LAGOSnutrient, by = "lagoslakeid") %>%
  select(lagoslakeid, sampledate, tn, tp, state, state_name) %>%
  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.

```
#violin plot comparing TN across states
TNstateviolin <- ggplot(LAGOSNandP, aes(x = state, y = tn)) +
  geom_violin(draw_quantiles = 0.50) +
  labs(x = "State", y = expression(Total ~ N ~ (mu*g / L)))
print(TNstateviolin)
```




```
#violin plot comparing TP across states
TPstateviolin <- ggplot(LAGOSNandP, aes(x = state, y = tp)) +
  geom_violin(draw_quantiles = 0.50) +
  labs(x = "State", y = expression(Total ~ P ~ (mu*g / L)))
print(TPstateviolin)
```



Which states have the highest and lowest median concentrations?

TN: Iowa has the highest median concentration and Vermont has the lowest median concentration.

TP: Illinois has the highest median concentration and Maine has the lowest median concentration.

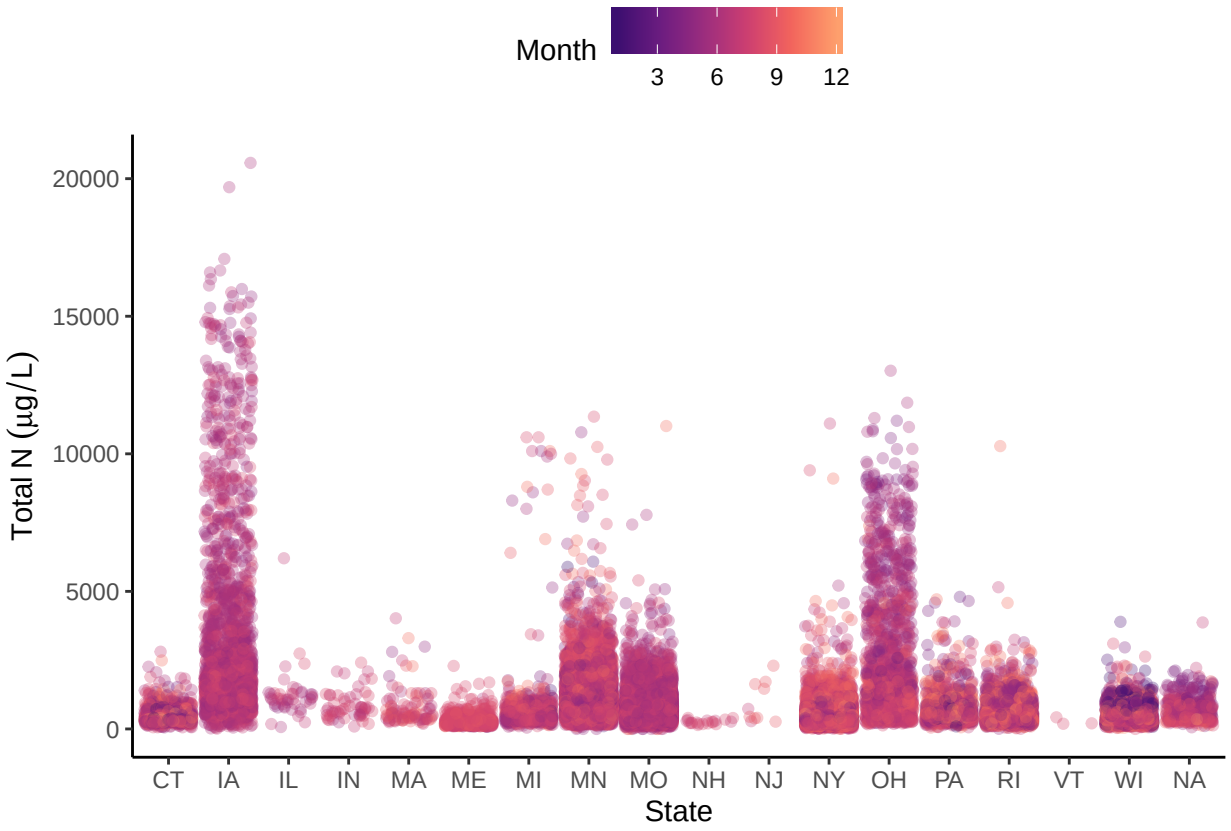
Which states have the highest and lowest concentration ranges?

TN: Iowa has the highest concentration range and New Hampshire and Vermont have the lowest concentration ranges.

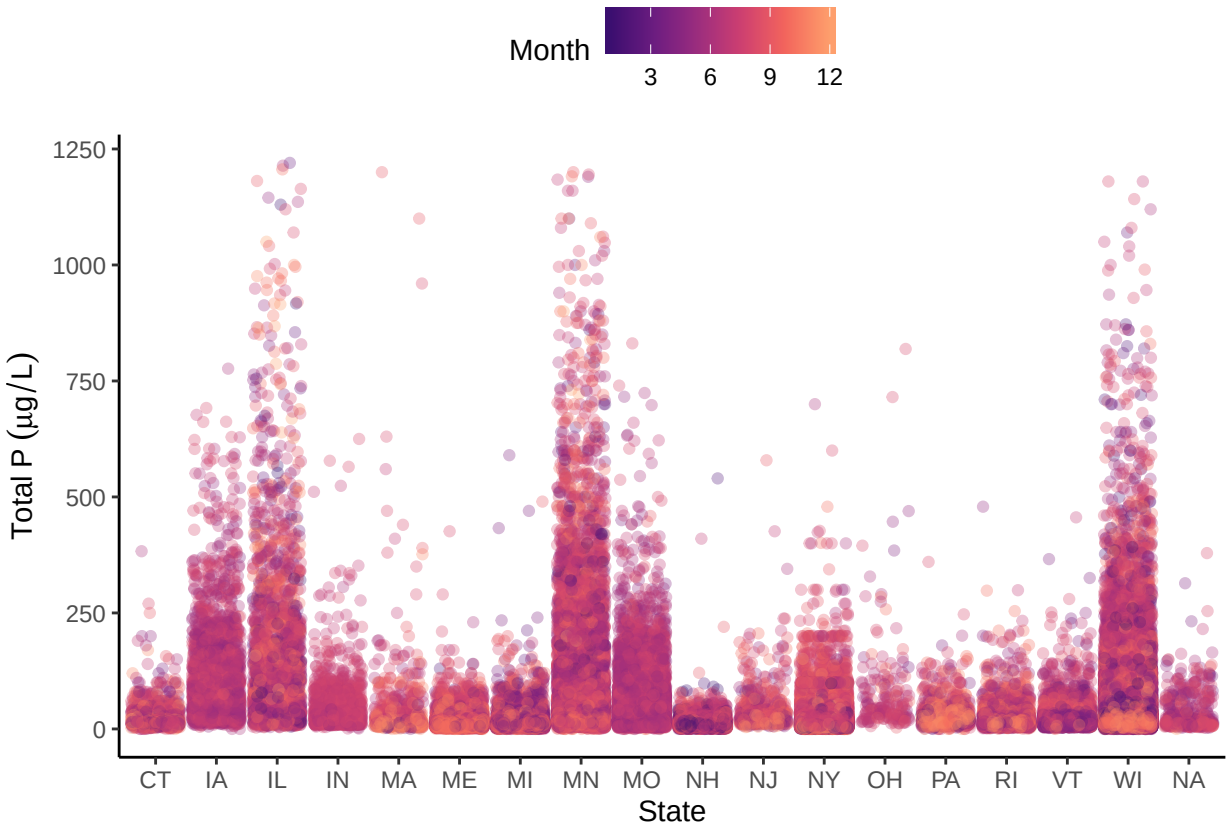
TP: Illinois and Minnesota have the highest concentration ranges and Connecticut and Pennslyvania have the lowest concentration ranges.

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.

```
#jitter ggplot of TN concentrations across states
TNstate_jitter <- ggplot(LAGOSNandP, aes(x = state, y = tn, color = samplemonth)) +
 geom_jitter(alpha = 0.3) +
 labs(x = "State", y = expression(Total ~ N ~ (mu*g / L)), color = "Month") +
 scale_color_viridis_c(option = "magma", begin = 0.2, end = 0.8) +
 theme(legend.position = "top")
print(TNstate_jitter)
```




```
#jitter ggplot of TP concentrations across states
TPstate_jitter <- ggplot(LAGOSNandP, aes(x = state, y = tp, color = samplemonth)) +
 geom_jitter(alpha = 0.3) +
 labs(x = "State", y = expression(Total ~ P ~ (mu*g / L)), color = "Month") +
 scale_color_viridis_c(option = "magma", begin = 0.2, end = 0.8) +
 theme(legend.position = "top")
print(TPstate_jitter)
```



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

TN: Iowa has the most samples, which is likely why Iowa had the highest median concentration and highest range.

TP: Illinois and Minnesota look like they have the most samples. This definitely impacted the highest median concentration (Illinois) and highest concentration ranges (Illinois and Minnesota).

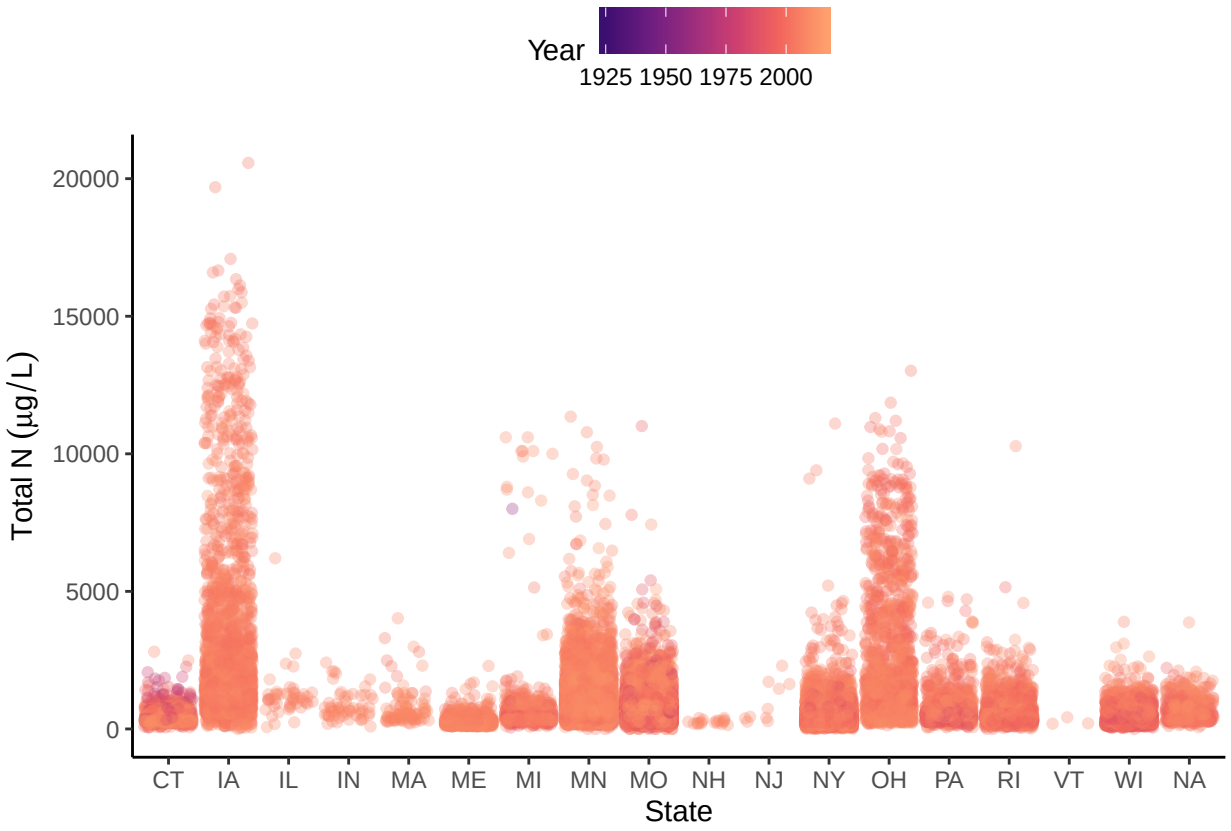
Which months are sampled most extensively? Does this differ among states?

TN: The summer to fall months are sampled most extensively. Yes, this does differ among states as some states have samples in the winter (i.e. Wisconsin).

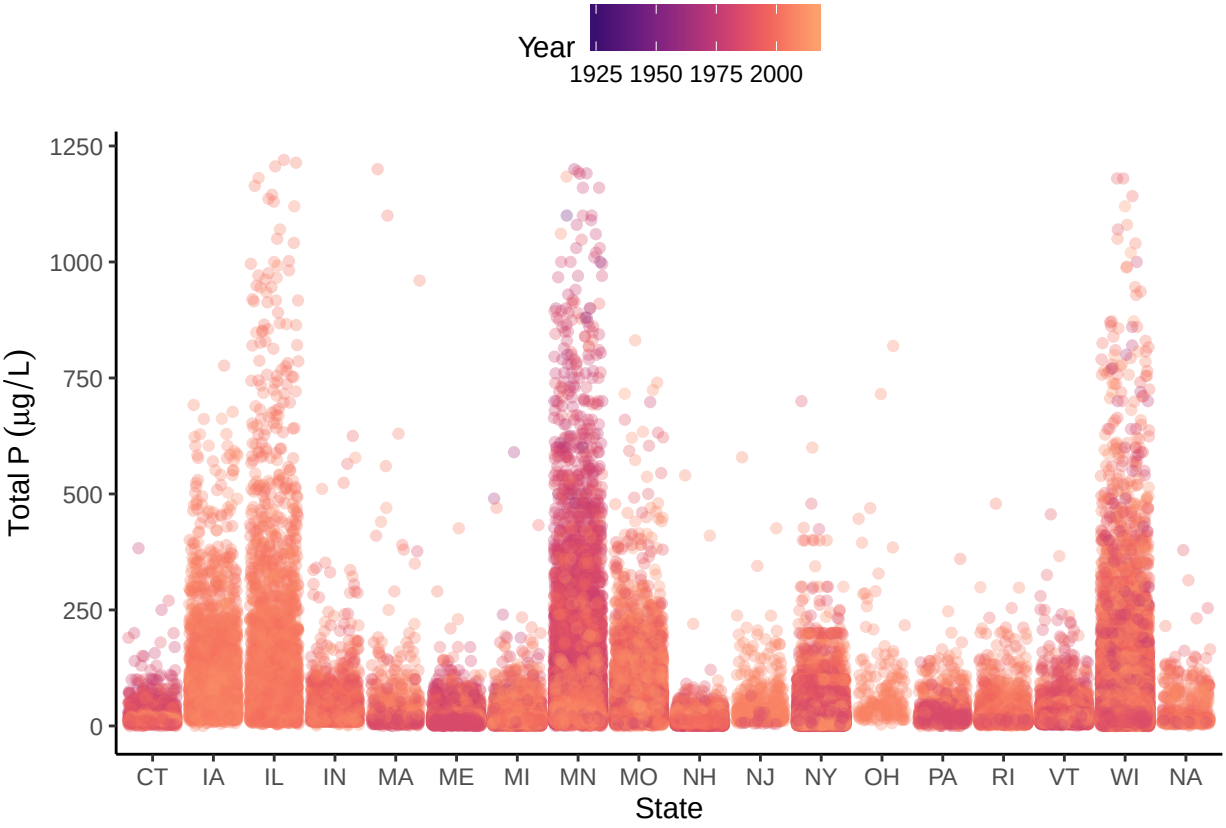
TP: The summer and fall monts look as if they are sampled most extensively. However, some states have samples from the winter months and fall months.

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.

```
#jitter plot of TN across state with sampleyear as color
TNstate_jitter2 <- ggplot(LAGOSNandP, aes(x = state, y = tn, color = sampleyear)) +
 geom_jitter(alpha = 0.3) +
 labs(x = "State", y = expression(Total ~ N ~ (mu*g / L)), color = "Year") +
 scale_color_viridis_c(option = "magma", begin = 0.2, end = 0.8) +
 theme(legend.position = "top")
print(TNstate_jitter2)
```




```
#jitter plot of TP across state with sampleyear as teh color
TPstate_jitter2 <- ggplot(LAGOSNandP, aes(x = state, y = tp, color = sampleyear)) +
 geom_jitter(alpha = 0.3) +
 labs(x = "State", y = expression(Total ~ P ~ (mu*g / L)), color = "Year") +
 scale_color_viridis_c(option = "magma", begin = 0.2, end = 0.8) +
 theme(legend.position = "top")
print(TPstate_jitter2)
```



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

TN: Years 1980s - 2016 were sampled most frequently. This does differ a little among states, with CT having some sampling occuring between 1950s - 1975.

TP: The years 1975 - 2016 were sampled most frequently for total phosphorus. However, this does differ among states as some states, such as MN, CT, ME, and VT have samples from the 1950s (or earlier) included in the dataset.

Reflection

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

    Spatial variety (i.e. where the states are located, some are further north than others) has an impact on lake water quality, particularly TN and TP. Tophic state calculated by secchi depth, TP, and/or TN has varying numbers of lakes in those categories.

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

    The jitter plot & violin plots help explain the variation and range of TN and TP in various states. To look at the number of lakes in each trophic state, the count function was used and helps support conclusions in 12.

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

    Yes, hands-on data analysis impacted my learning about water quality relative to theory-based because it provided me with visualization as well as a better understanding of the data we were working with.

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

Real-world data is always "messier" than the theory data. The data here did follow along with what we would expect to find in water quality of lakes. There were some columns that had NAs and those were obviously removed when plots were created.