Assignment 5: Water Quality in Lakes

Keqi He

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 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., "A05_Salk.pdf") 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] "C:/Users/kh382/Documents/Hydrologic_Data_Analysis"
```

```
#install.packages("tidyverse")
#install.packages("lubridate")
#install.packages("LAGOSNE")

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

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## 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 ------------------------------------------------------------------------------------ tidy
## 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())

load(file = "./Data/Raw/LAGOSdata.rda")
LAGOStrophic <- 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).

```
LAGOStrophic <-
   mutate(LAGOStrophic,
          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.

```
LAGOStrophic %>% count(trophic.class)
```

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

```
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
```

```
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 eutrophic or hypereutrophic according to the three different metrics (trophic.class, trophic.class.secchi, trophic.class.tp)?

```
mean(LAGOStrophic$trophic.class == "Eutrophic")
```

```
## [1] 0.5585116
```

```
mean(LAGOStrophic$trophic.class == "Hypereutrophic")
```

```
## [1] 0.1918453
```

```
mean(LAGOStrophic$trophic.class.secchi == "Eutrophic")
```

```
## [1] 0.3823698
```

```
mean(LAGOStrophic$trophic.class.secchi == "Hypereutrophic")
```

```
## [1] 0.06803111
```

```
mean(LAGOStrophic$trophic.class.tp == "Eutrophic")
```

```
## [1] 0.3314032
```

```
mean(LAGOStrophic$trophic.class.tp == "Hypereutrophic")
```

```
## [1] 0.09643634
```

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

> Secchi disk transparency. Among three different metrics (trophic.class, trophic.class.secchi, trophic.class.tp), proportion of total observations considered eutrophic or hypereutrophic is smallest. Reasons: some dissolved organic matters, phosphorus or nitrogens that are colorless and transparent do not influence secchi disk transparenvy and therefore secchi disk transparenvy is most conservative in its designation of eutrophic conditions

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.

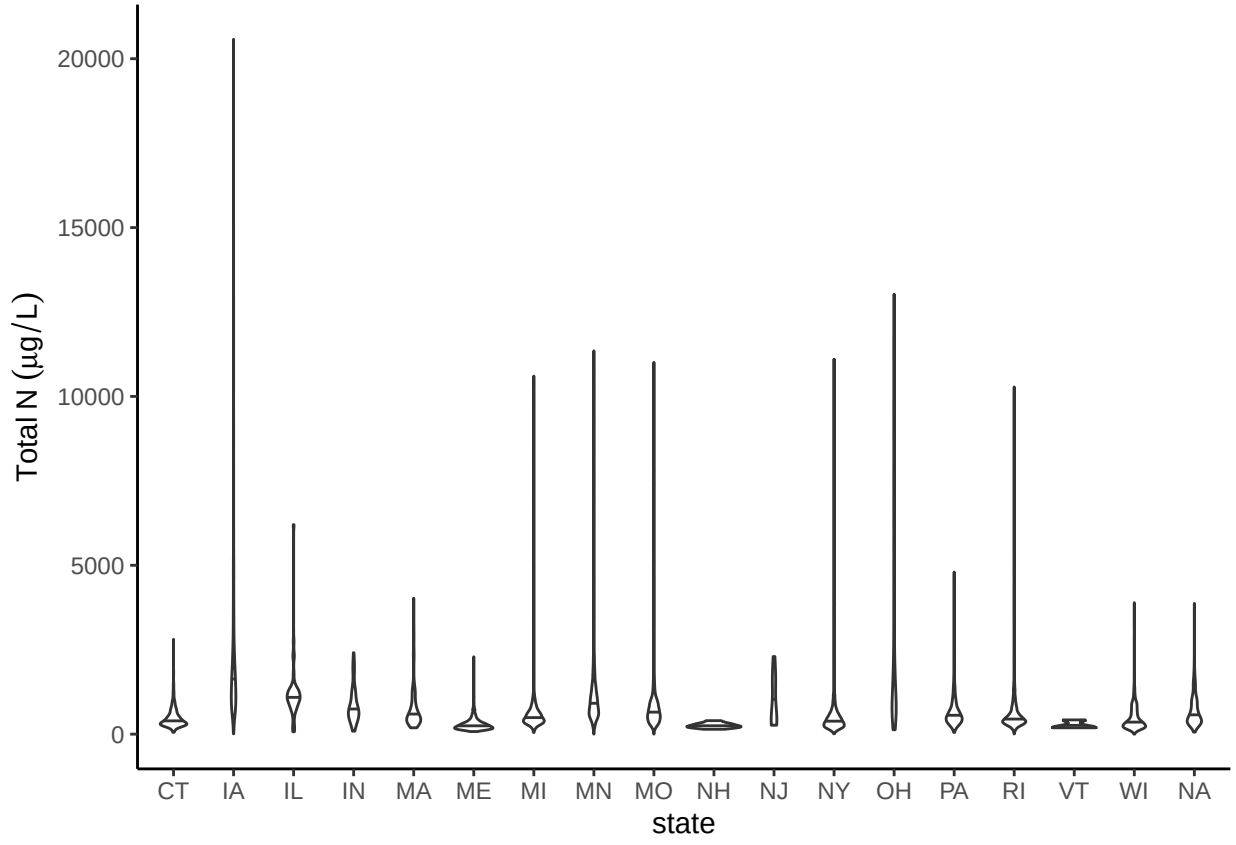
```
LAGOSlocus <- LAGOSdata$locus
LAGOSstate <- LAGOSdata$state
LAGOSnutrient <- LAGOSdata$epi_nutr

LAGOSlocations <- left_join(LAGOSlocus, LAGOSstate, by = "state_zoneid")
LAGOSNandP <-
   left_join(LAGOSnutrient, LAGOSlocations, 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.

```
LAGOSTNviolin <- ggplot(LAGOSNandP, aes(x = state, y = tn)) +
  geom_violin(draw_quantiles = 0.50) +
  labs(y = expression(Total ~ N ~ (mu*g / L)))
print(LAGOSTNviolin)
```

```
## Warning: Removed 774226 rows containing non-finite values (stat_ydensity).
```

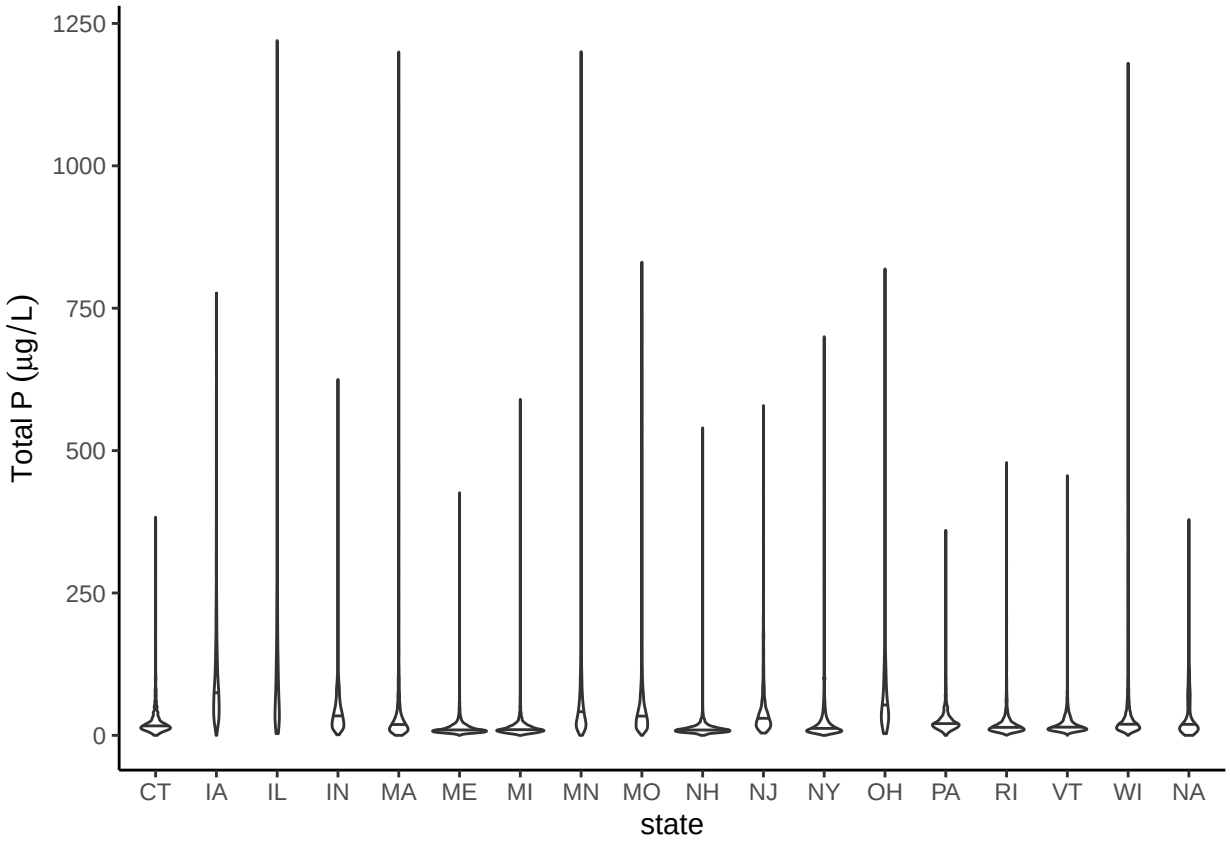
```
## Warning in regularize.values(x, y, ties, missing(ties)): collapsing to
## unique 'x' values

## 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

## 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
```




```
LAGOSTPviolin <- ggplot(LAGOSNandP, aes(x = state, y = tp)) +
  geom_violin(draw_quantiles = 0.50) +
  labs(y = expression(Total ~ P ~ (mu*g / L)))
print(LAGOSTPviolin)
```

```
## Warning: Removed 672861 rows containing non-finite values (stat_ydensity).
```

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



Which states have the highest and lowest median concentrations?

TN: IA (highest), VT (lowest)

TP: IL (highest), ME (lowest)

Which states have the highest and lowest concentration ranges?

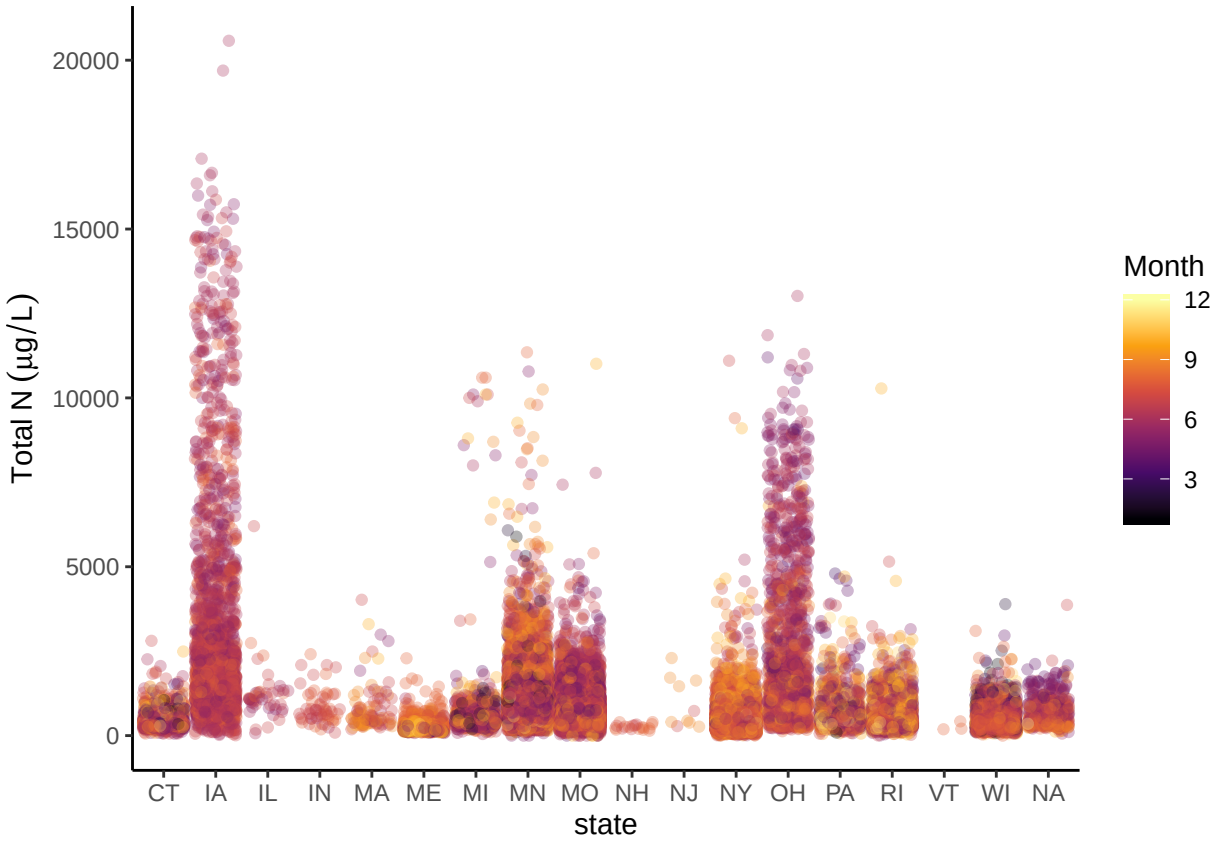
TN: IA (highest), VT (lowest)

TP: IL (highest), PA (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.

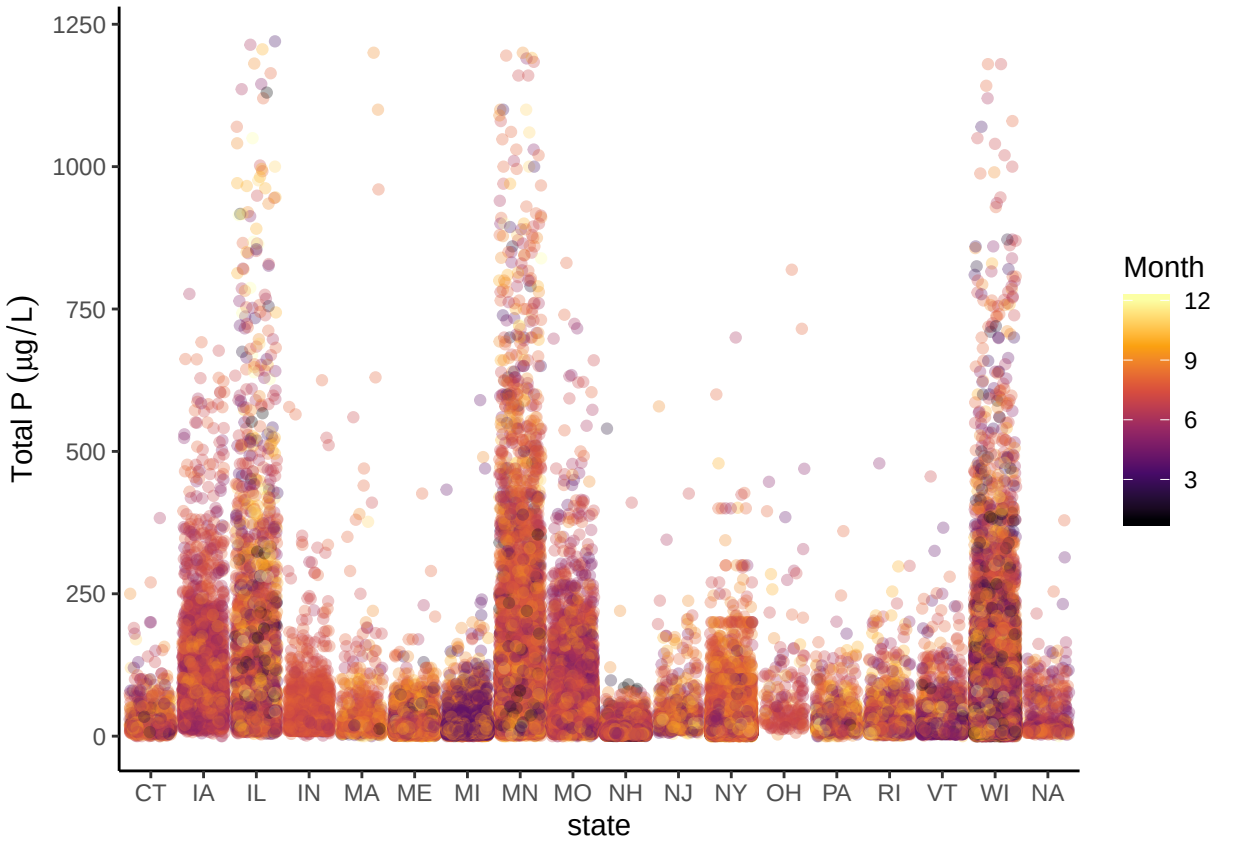
```
stateTNjitter <- ggplot(LAGOSNandP, aes(x = state, y = tn, color = samplemonth)) +
  geom_jitter(alpha = 0.3) +
  labs(y = expression(Total ~ N ~ (mu*g / L)), color = "Month") +
  scale_color_viridis_c(option = "inferno")
print(stateTNjitter)
```

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




```
stateTPjitter <- ggplot(LAGOSNandP, aes(x = state, y = tp, color = samplemonth)) +
  geom_jitter(alpha = 0.3) +
  labs(y = expression(Total ~ P ~ (mu*g / L)), color = "Month") +
  scale_color_viridis_c(option = "inferno")
print(stateTPjitter)
```

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



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

TN: MO

TP: WI

The range will increase accordingly as the number of samples increases. But, it does not mean the most samples lead to biggest total range. More sample size makes the range more convincing and representative.

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

TN: July

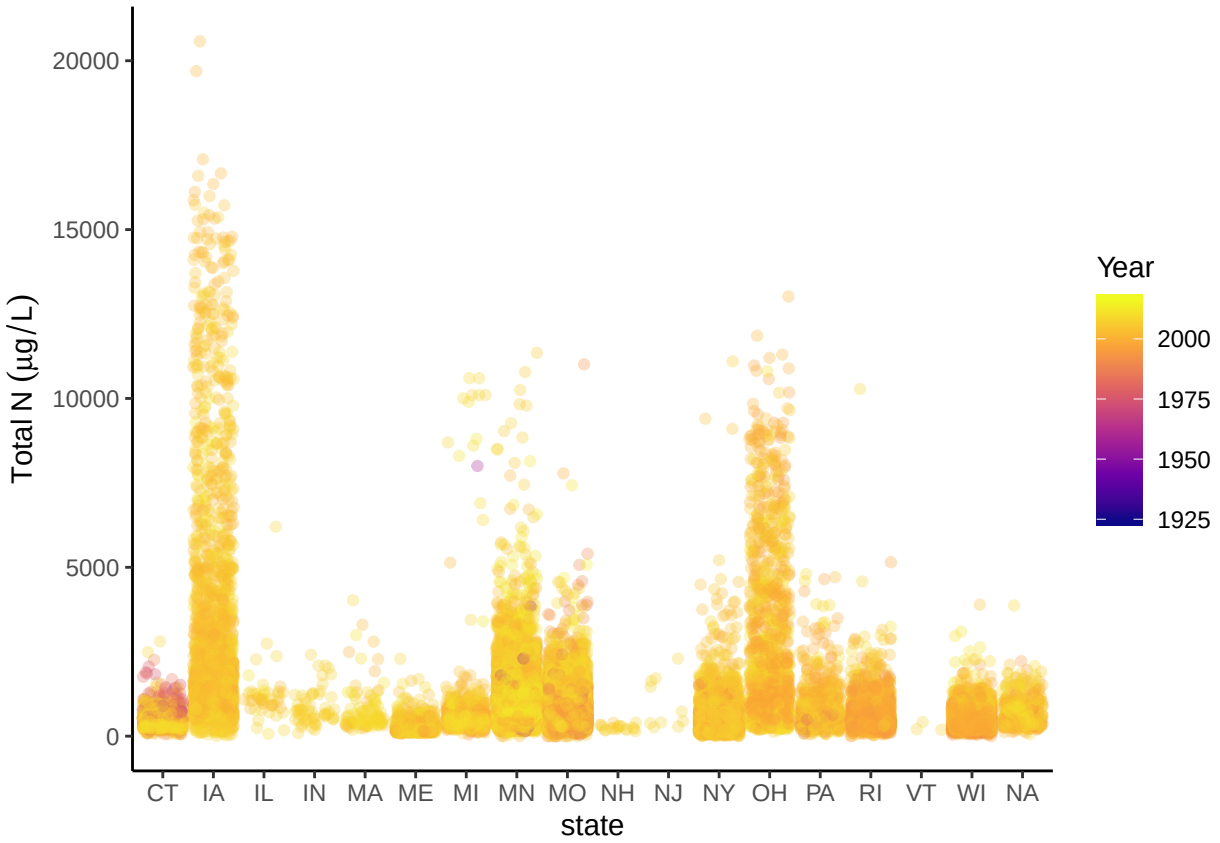
TP: August

Yes, this differs among states. Like ME sampled total nitrogens most extensively in August. But almost every state sampled total nitrogens and total phosphorus most extensively in summer (mainly 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.

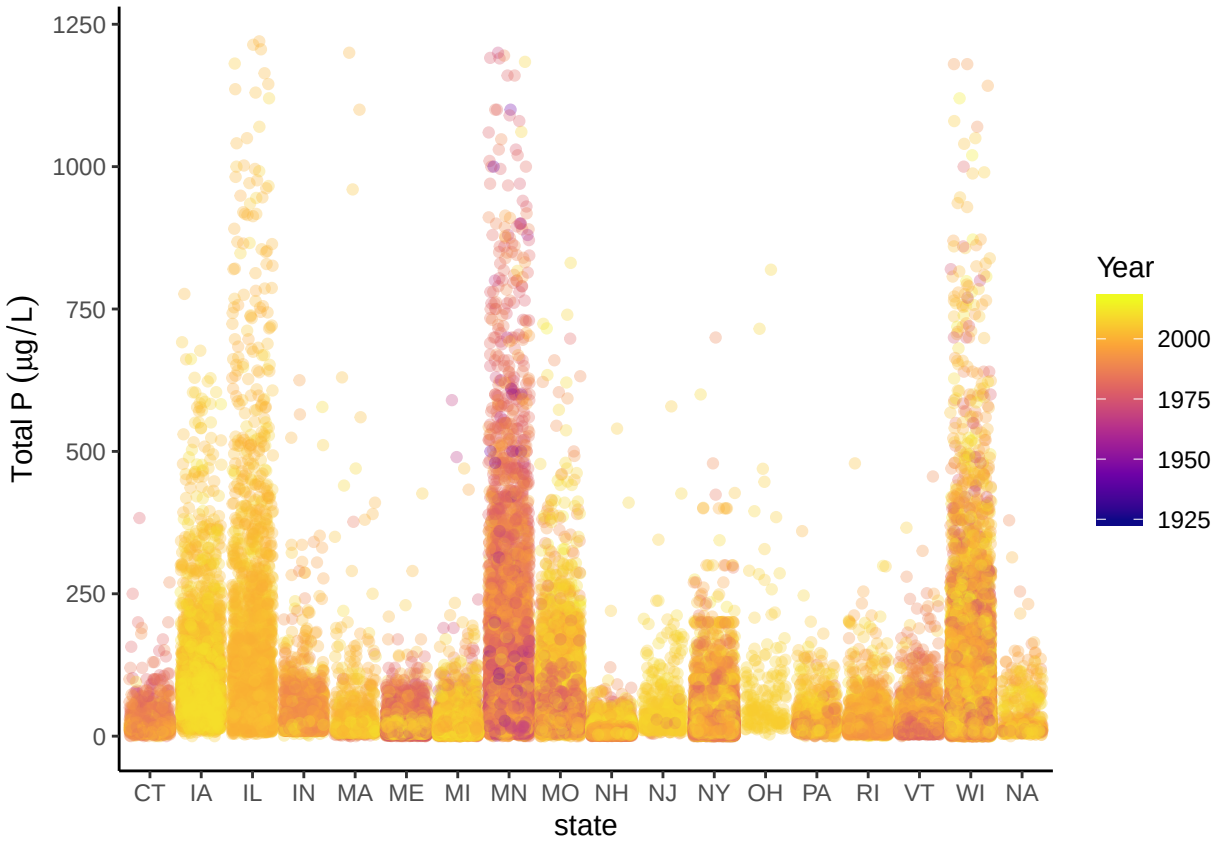
```
stateTNjitteryear <- ggplot(LAGOSNandP, aes(x = state, y = tn, color = sampleyear)) +
  geom_jitter(alpha = 0.3) +
  labs(y = expression(Total ~ N ~ (mu*g / L)), color = "Year") +
  scale_color_viridis_c(option = "plasma")
print(stateTNjitteryear)
```

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




```
stateTPjitteryear <- ggplot(LAGOSNandP, aes(x = state, y = tp, color = sampleyear)) +
  geom_jitter(alpha = 0.3) +
  labs(y = expression(Total ~ P ~ (mu*g / L)), color = "Year") +
  scale_color_viridis_c(option = "plasma")
print(stateTPjitteryear)
```

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



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

TN: 2009

TP: 2009

Yes, this differs among states. Like ME sampled total phosphorus most extensively in 1976.

Reflection

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

    Dissolved oxygen, nutrients, microbes, odor/ color, invasive species, contaminants, etc are major water quality impairments experienced in lakes. Trophic state given by Robert Carlson in 1977 is a useful water quality metric which means the amount of biomass a given system can sustain. To calculate the Trophic State Index, three variables can be used: chlorophyll a concentration, Secchi disk transparency and Total phosphorus (TP).

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

    Data used includes LAGOSdata database and the trophic state index data. Ggplot, Violin plots and jitter plots are used to draw those conclusions.

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 can help me learn lake water quality greatly. It is more vivid and memorable and can help me understand the theory better.

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

Real-world data is complicated and imperfect because there are lots of factors can affect the value of real-world data.Different states have different measurement extensity. But in general, real-world data has a theoretical law.