Assignment 5: Water Quality in Lakes

Lindsay Roth

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/lindsayroth/Documents/MEM 2nd Year/HydroData/Hydrologic_Data_Analysis"
```

```
library(tidyverse)
```

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

```
## 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_co
## 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())
```

```
LAGOSdata <- lagosne_load()
```

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

```
TrophicData <- 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).

```
TrophicData <- TrophicData %>%
  mutate(trophic.class.secchi =
           ifelse(TSI.secchi < 40, "Oligotrophic",
                  ifelse(TSI.secchi < 50, "Mesotrophic",
                         ifelse(TSI.secchi < 70, "Eutrophic", "Hypereutrophic")))) %>%
  mutate(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.

```
count(TrophicData, trophic.class)
```

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

```
count(TrophicData, 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
```

```
count(TrophicData, 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)?

```
#trophic.class
##Hypereutrophic
14379/74951
```

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

```
##0.19 or 19%

##Eutrophic
41861/74951
```

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

```
##0.56 or 56%


#trophic.class.secchi
##Hypereutrophic
5099/74951
```

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

```
##0.07 or 7%

##Eutrophic
28659/74951
```

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

```
##0.38 or 38%


#trophic.class.tp
##Hypereutrophic
7228/74951
```

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

```
##0.096 or 9.6%

##Eutrophic
24839/74951
```

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

```
##0.33 or 33%
```

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

> Total phosphorus seems to be the most conservativie in designating eutrophic conditions. This may be because there are other factors influencing higher levels of chlorophyll a and secchi depth, such as total nitrogen, turbidity, sediment load, etc. Total phosphorus levels are not influenced by any of these factors.

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

LAGOSlocus$lagoslakeid <- as.factor(LAGOSlocus$lagoslakeid)
LAGOSnutrient$lagoslakeid <- as.factor(LAGOSnutrient$lagoslakeid)

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

LAGOSlocations <-
  within(LAGOSlocations,
         state <- factor(state, levels = names(sort(table(state), decreasing=TRUE))))

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

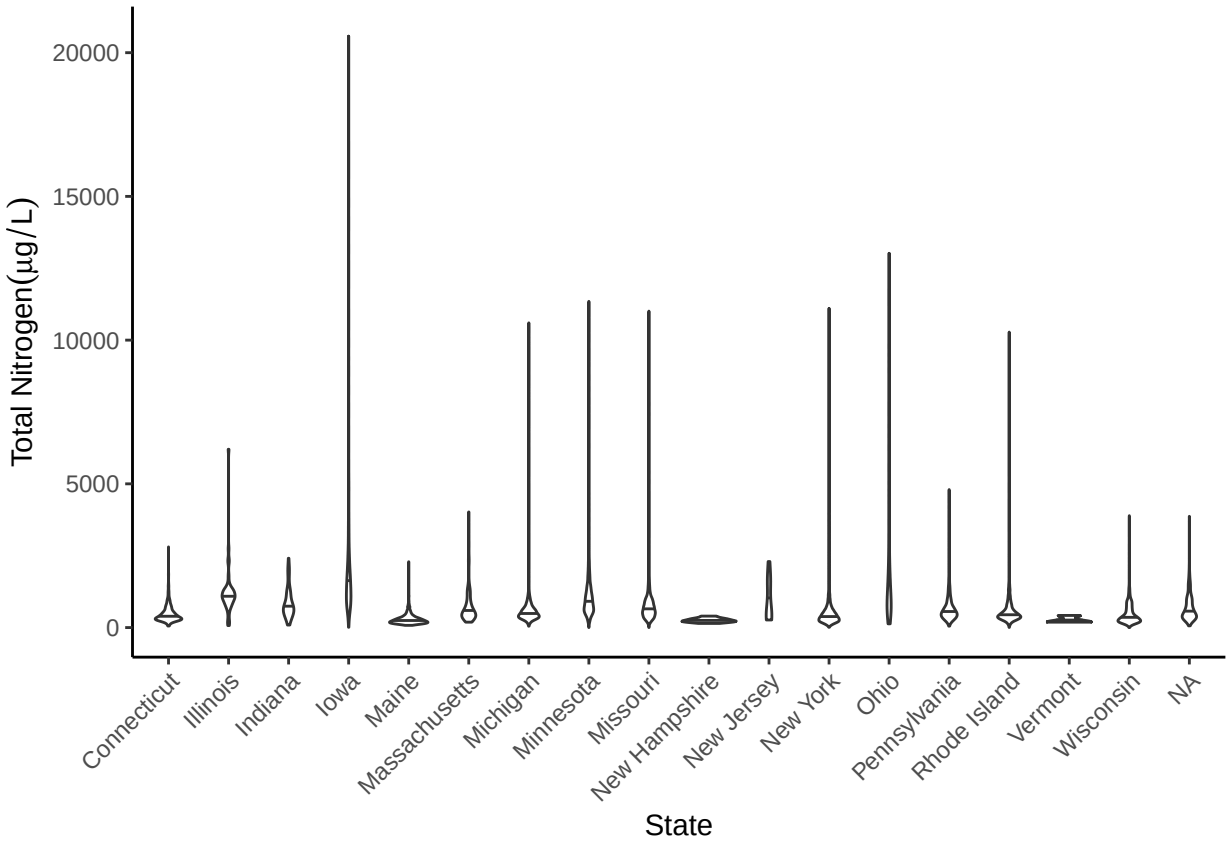
```
## Warning: Column `lagoslakeid` joining factors with different levels,
## coercing to character vector
```

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

```
NPlot <- ggplot(LAGOSNandP) +
  geom_violin(aes(x = state_name, y = tn), draw_quantiles = 0.5) +
  labs(x = "State", y =  expression("Total Nitrogen" (mu*g / L))) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
print(NPlot)
```

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

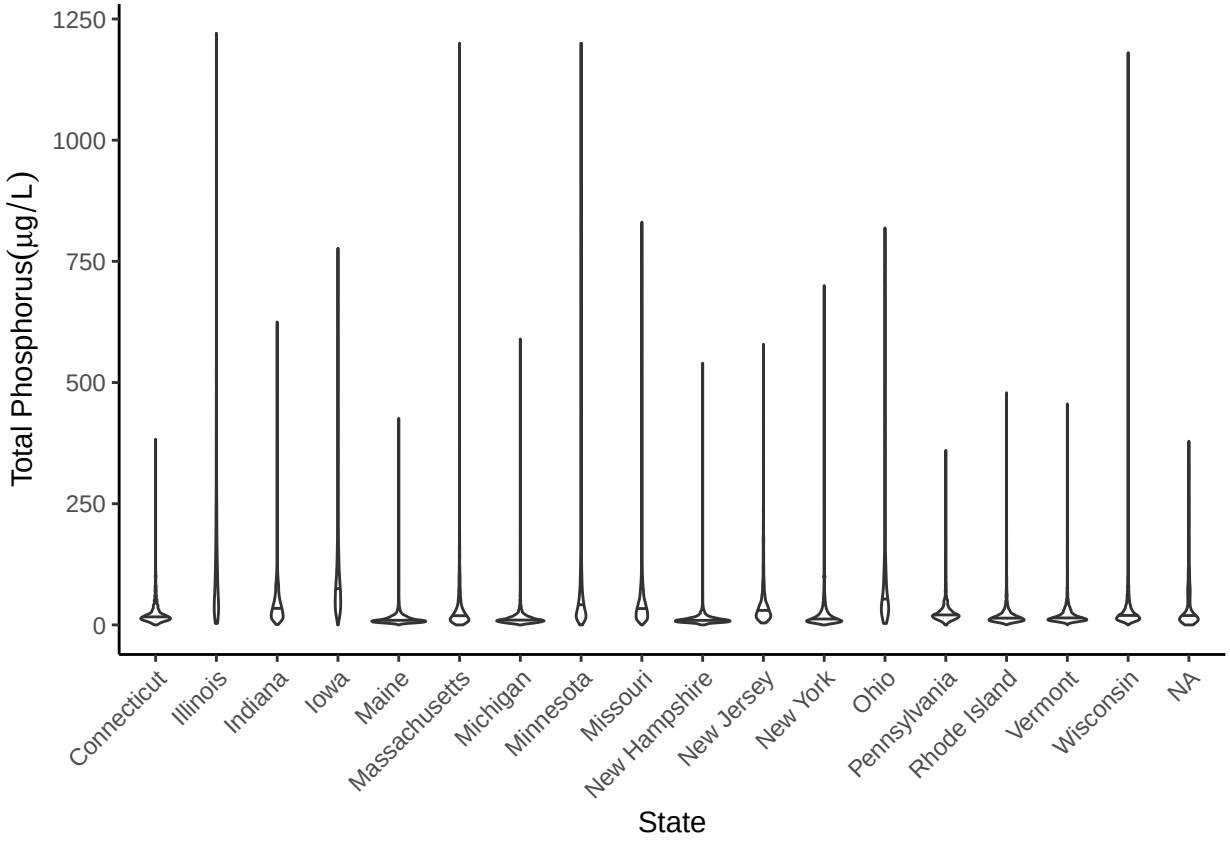
```
## Warning in regularize.values(x, y, ties, missing(ties)): collapsing to
## unique 'x' values

## 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
```




```
PPlot <- ggplot(LAGOSNandP) +
  geom_violin(aes(x = state_name, y = tp), draw_quantiles = 0.5) +
  labs(x = "State", y =  expression("Total Phosphorus" (mu*g / L))) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
print(PPlot)
```

```
## 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

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



Which states have the highest and lowest median concentrations?

TN: Iowa and Ohio have the highest, New Hampshire Maine and Vermont have the lowest

TP: Iowa and Illinois have the highest, Maine Michigan and New Hampshire have the lowest

Which states have the highest and lowest concentration ranges?

TN: Iowa has the highest range, New Hampshire has the lowest range

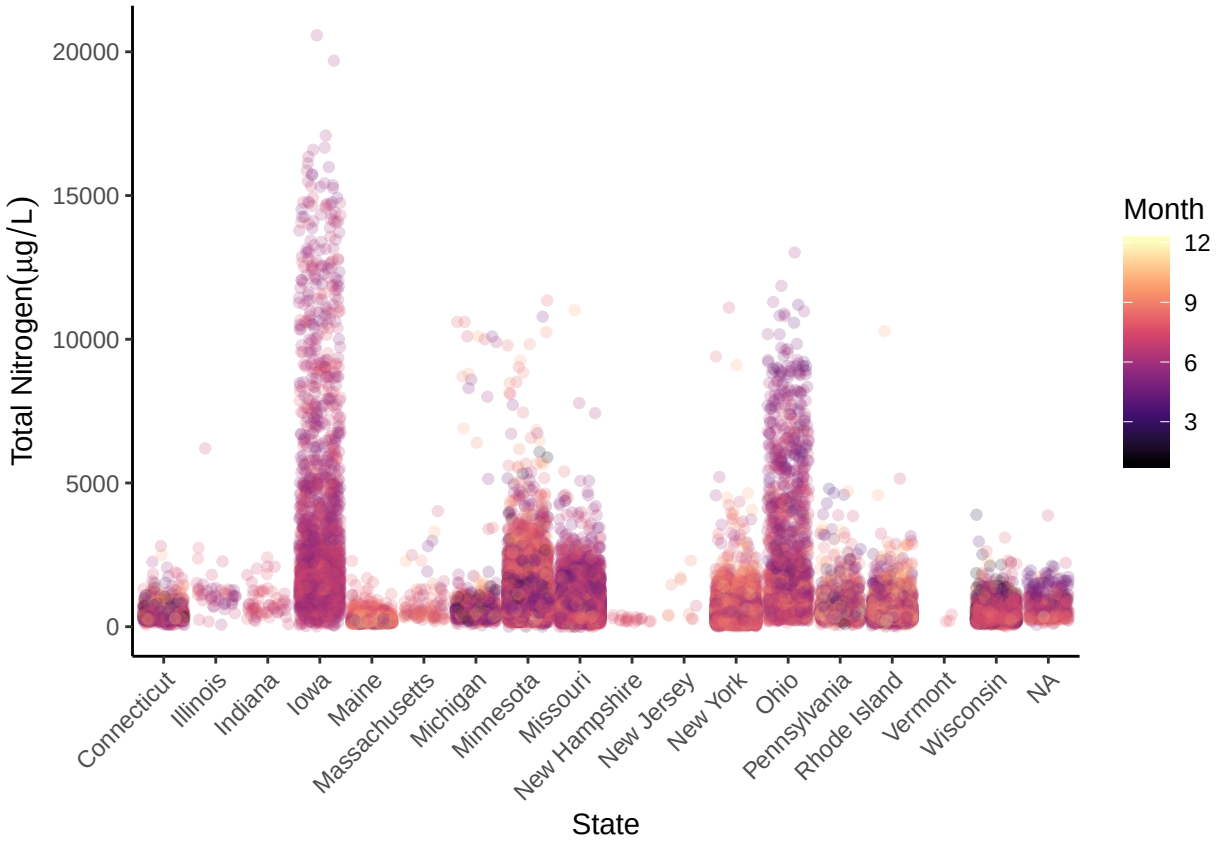
TP: Illinois has the highest range, Pennsylvania has the lowest range

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.

```
N.by.Month <-
ggplot(LAGOSNandP,
       aes(x = as.factor(state_name), y = tn, color = samplemonth)) +
  geom_jitter(alpha = 0.2) +
  labs(x = "State", y = expression("Total Nitrogen"(mu*g / L)), color = "Month") +
  scale_color_viridis_c(option = "magma") +
```

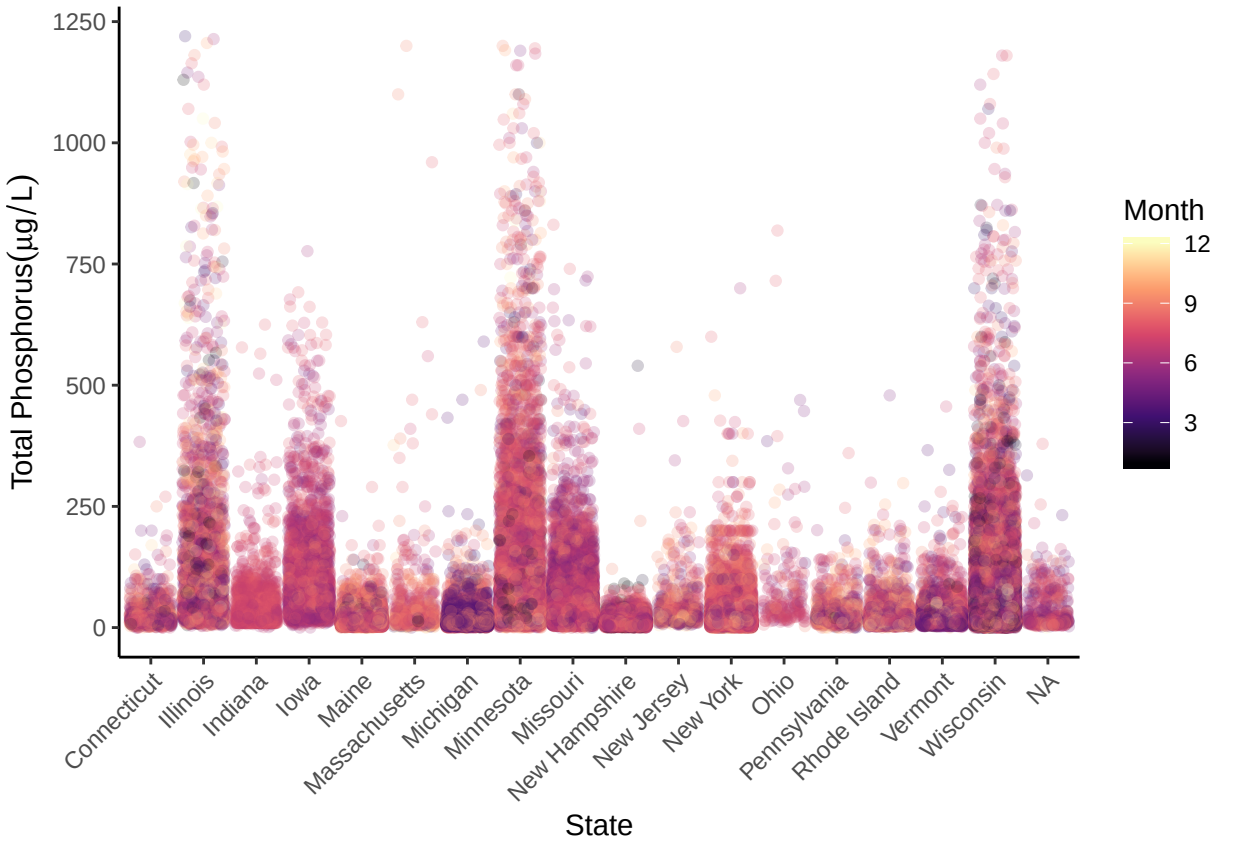
```
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
print(N.by.Month)
```

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




```
P.by.Month <-
ggplot(LAGOSNandP,
       aes(x = as.factor(state_name), y = tp, color = samplemonth)) +
  geom_jitter(alpha = 0.2) +
  labs(x = "State", y = expression("Total Phosphorus"(mu*g / L)), color = "Month") +
  scale_color_viridis_c(option = "magma") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
print(P.by.Month)
```

```
## 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: It looks like Iowa, Ohio, Minnesota, and Missouri have taken the most tn samples. This may have contributed to the wide range of values for Iowa and Ohio.
>
> TP: Illinois, Iowa, Minnesota, Missouri, and Wisconsive have taken the most tp samples. This may have contributed to the wide range of values for Illinois, Minnesota, and Wisconsin.

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

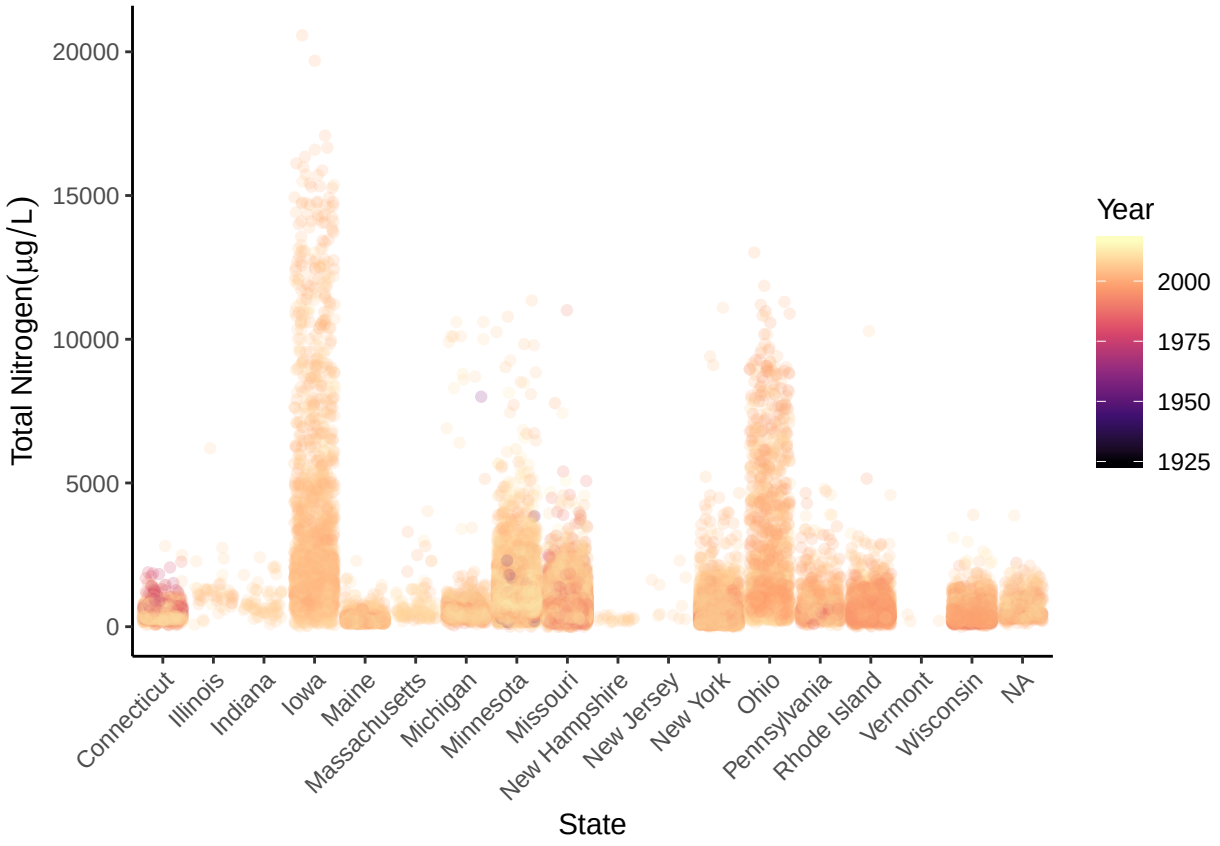
> TN: It looks like May, June, July, and August were most extensively sampled. Some states like Indiana appear to have only sampled during one month (likely July), while other states have more samples later in the year, like New York and Rhode Island as well as earlier in the year like Michigan and Ohio. Not many samples in any states were taken in the winter months.
>
> TP: Phosphorus is similar to Nitrogen. THe most extensively sampled months were June, July, and August, with Indiana appearing to only sample in July and August while Michigan, Vermont, and Wisconsin have samples from earlier and later in the year. Not many samples in any states were taken in the winter 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.

```
N.by.Year <-
ggplot(LAGOSNandP,
       aes(x = as.factor(state_name), y = tn, color = sampleyear)) +
  geom_jitter(alpha = 0.2) +
  labs(x = "State", y = expression("Total Nitrogen"(mu*g / L)), color = "Year") +
  scale_color_viridis_c(option = "magma") +
```

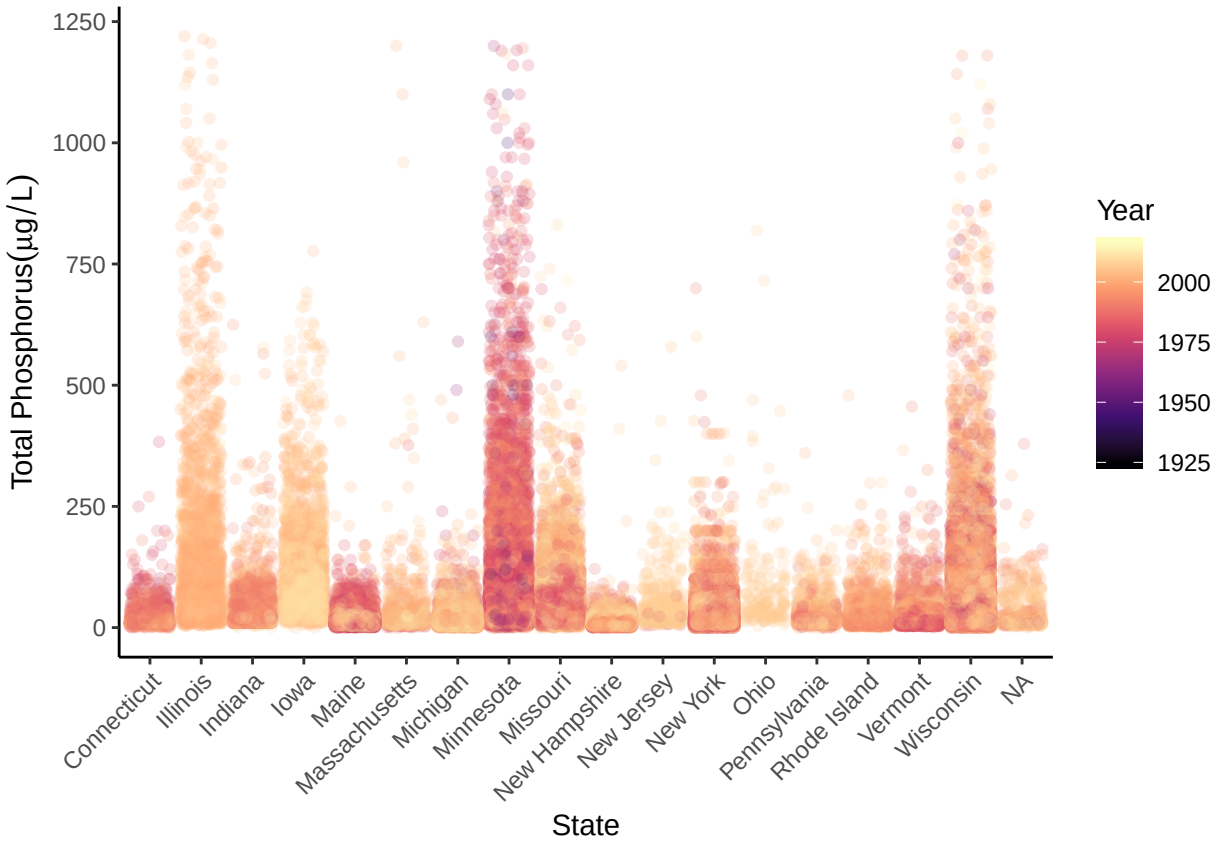
```
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
print(N.by.Year)
```

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




```
P.by.Year <-
ggplot(LAGOSNandP,
       aes(x = as.factor(state_name), y = tp, color = sampleyear)) +
  geom_jitter(alpha = 0.2) +
  labs(x = "State", y = expression("Total Phosphorus"(mu*g / L)), color = "Year") +
  scale_color_viridis_c(option = "magma") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
print(P.by.Year)
```

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



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

TN: The years that were sampled most extensively include the late 1990s and 2000s. Connecticut has some mesurements from what appears to be the 1970s and 1980s.

TP: The Years that were sampled most were the 1990s and the 2000s. Minnesota also has samples starting from what looks like the 1950s. Connecticut, Maine, New York, Vermont, and Wisconsin have samples from the 1980s.

Reflection

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

    Conclusion 1: States with large agricultural economies have higher measurements of total nitrogen and total phosphorus than states with smaller agricultural economies. Coclusion 2: Total phosphorus, secchi depth, and chlorophyll a all predict slightly varied eutrophic levels in water bodies.

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

    In the jitter plots, the highest measurements of total nitrogen and total phosphorus were seen in Illinois, Indiana, Minnesota, Missouri, Ohio and Wisconsin.

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

    Yes, being able to visualize trends helps me learn concepts much more than just reading explanations.

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

    The data we used for this assignment aligned fairly well with my expectations. I know that a lot of nitrogen and phosphorus runs off from agricultural land due to excess fertilizers applied on the surface of the soil. These nutrients, when added to lakes, rivers, and estuaries,contribute to higher biological activity in these water bodies. Therefore, I expected states with higher agricultural activity to have higher trophic levels.