# Assignment 5: Water Quality in Lakes

*Tristen Townsend*

## 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/Tristen/OneDrive - Duke University/Fall 2019/Hydrologic Data Analysis/Hydrologic_Data_Ana
```

```
library(tidyverse)
```

```
## -- 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   0.8.3     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()
```

```
library(lubridate)
```

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

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

```
library(LAGOSNE)

theme_set(theme_classic())
```

```
options(scipen = 100)

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

```
## Parsed with column specification:
## cols(
##   lagoslakeid = col_double(),
##   sampledate = col_date(format = ""),
##   chla = col_double(),
##   tp = col_double(),
##   secchi = col_double(),
##   gnis_name = col_character(),
##   lake_area_ha = col_double(),
##   state = col_character(),
##   state_name = col_character(),
##   sampleyear = col_double(),
##   samplemonth = col_double(),
##   season = col_character(),
##   TSI.chl = col_double(),
##   TSI.secchi = col_double(),
##   TSI.tp = col_double(),
##   trophic.class = col_character()
## )
```

## 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,
         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.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"))))

LAGOStrophic$trophic.class <-
  factor(LAGOStrophic$trophic.class,
         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"))
```

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.class <- count(LAGOStrophic, trophic.class)
count.secchi <- count(LAGOStrophic, trophic.class.secchi)
count.tp <- count(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)?

```
#Trophic class - Chl.A
#Eutrophic
count.class[3,2]/sum(count.class$n) #0.559
```

```
##           n
## 1 0.5585116
```

```
#Hypereutrophic
count.class[4,2]/sum(count.class$n) #0.192
```

```
##           n
## 1 0.1918453
```

```
#Trophic class - Secchi
#Eutrophic
count.secchi[3,2]/sum(count.secchi$n) #0.382
```

```
##           n
## 1 0.3823698
```

```
#Hypereutrophic
count.secchi[4,2]/sum(count.secchi$n) #0.068
```

```
##            n
## 1 0.06803111
```

```
#Trophic class - TP
#Eutrophic
count.tp[3,2]/sum(count.tp$n) #0.331
```

```
##           n
## 1 0.3314032
```

```
#Hypereutrophic
count.tp[4,2]/sum(count.tp$n) #0.096
```

```
##            n
## 1 0.09643634
```

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

> Total phosphorus. This is likely because there are more things that could influence the other two variables. For secchi depth, it can be affected by things such as dissolved organics or sediments in the water. And chorophyll-a has more things that can influence it's values that do not affect total phosphorus concentrations (for example, nitrogen would might increase chlorophyll-a but not total phosphorus).

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")

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

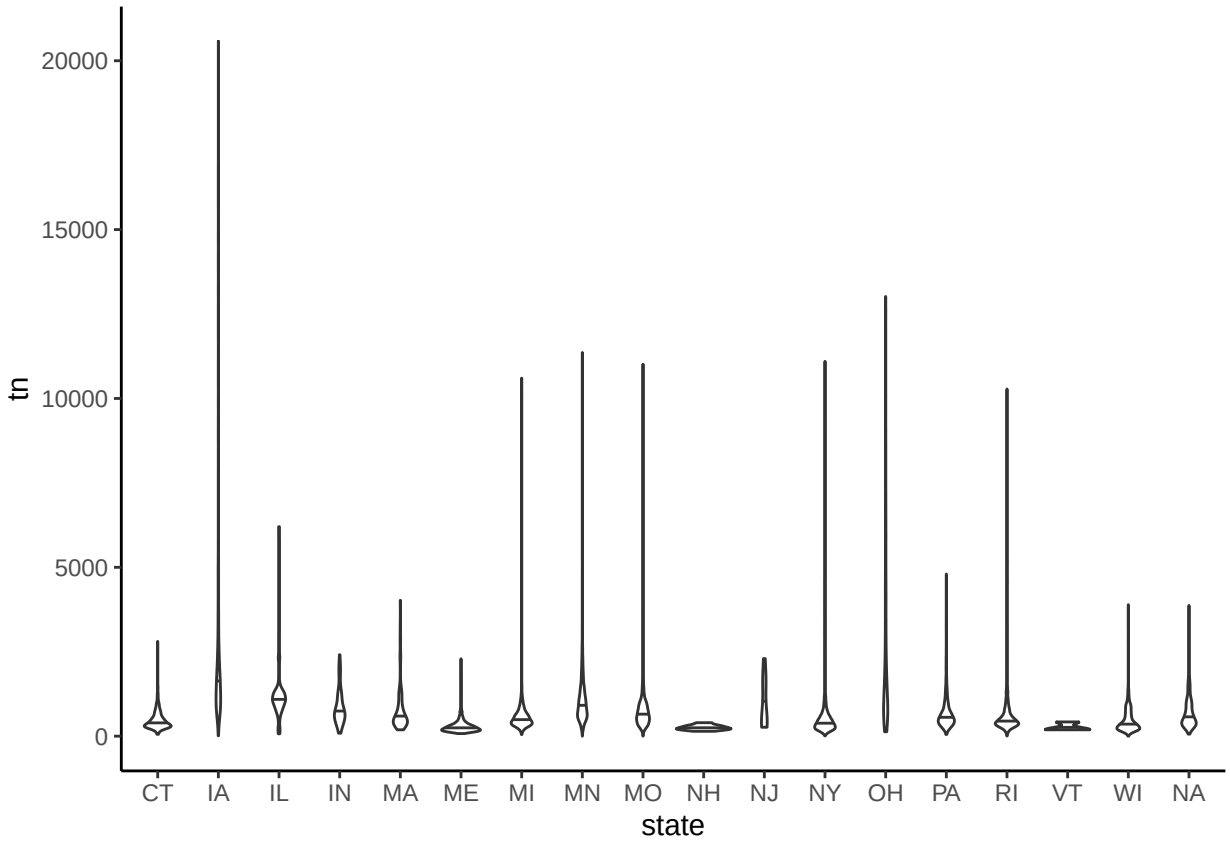
```
## Warning: Column `lagoslakeid` joining factors with different levels,
## coercing to character vector
```

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

```
stateTNviolin <- ggplot(LAGOSNandP, aes(x = state, y = tn)) +
  geom_violin(draw_quantiles = 0.50)
print(stateTNviolin)
```

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

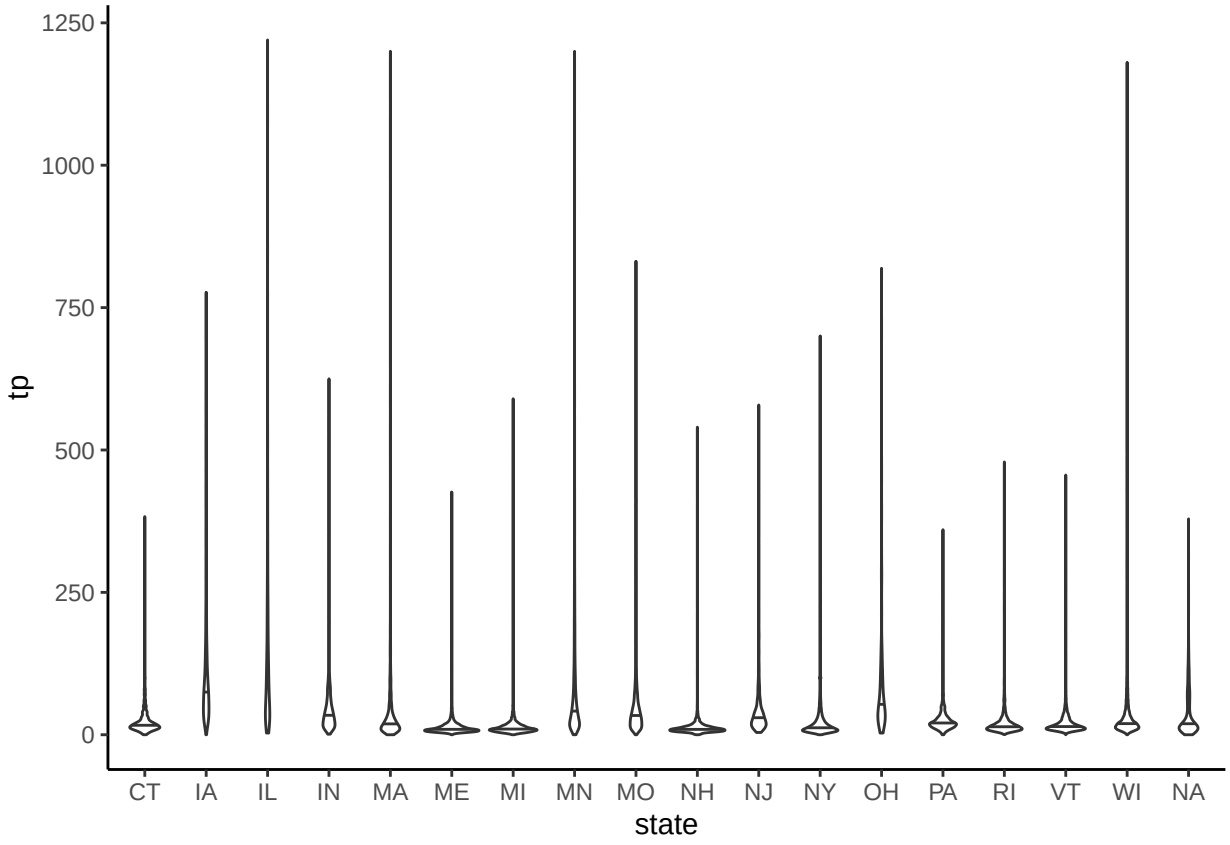
```
## Warning in regularize.values(x, y, ties, missing(ties)): collapsing to
## unique 'x' values

## 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
```




```
stateTPviolin <- ggplot(LAGOSNandP, aes(x = state, y = tp)) +
  geom_violin(draw_quantiles = 0.50)
print(stateTPviolin)
```

```
## 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: Highest: Iowa; Lowest: Vermont, New Hamsphire, and Maine (this visualization makes it difficult to be certain)

TP: Highest: Illinois; Lowest: Maine (though this visualization makes it difficult to be certain)

Which states have the highest and lowest concentration ranges?

TN: Highest range: Iowa; Lowest range: Vermont or New Hampshire

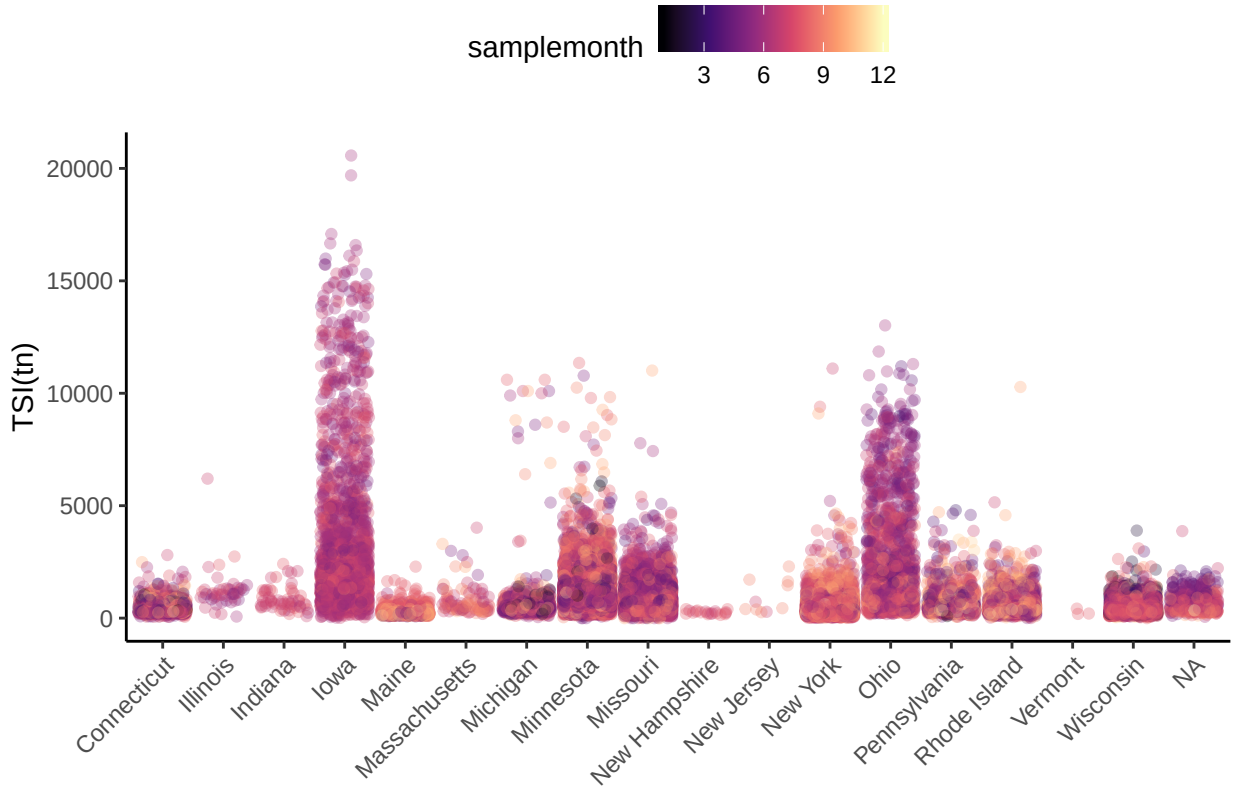
TP: Highest range: Illinois, Massachusetts, Minnesota; Lowest range: Pennsylvania

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.

```
LAGOSNandP$state <- as.factor(LAGOSNandP$state)
LAGOSNandP$state_name <- as.factor(LAGOSNandP$state_name)

stateTNjitter <- ggplot(LAGOSNandP, aes(x = state_name, y = tn, color = samplemonth)) +
  geom_jitter(alpha = 0.3) +
  labs(x = "", y = "TSI(tn)") +
  theme(legend.position = "top") +
  scale_color_viridis_c(option = "magma") +
```

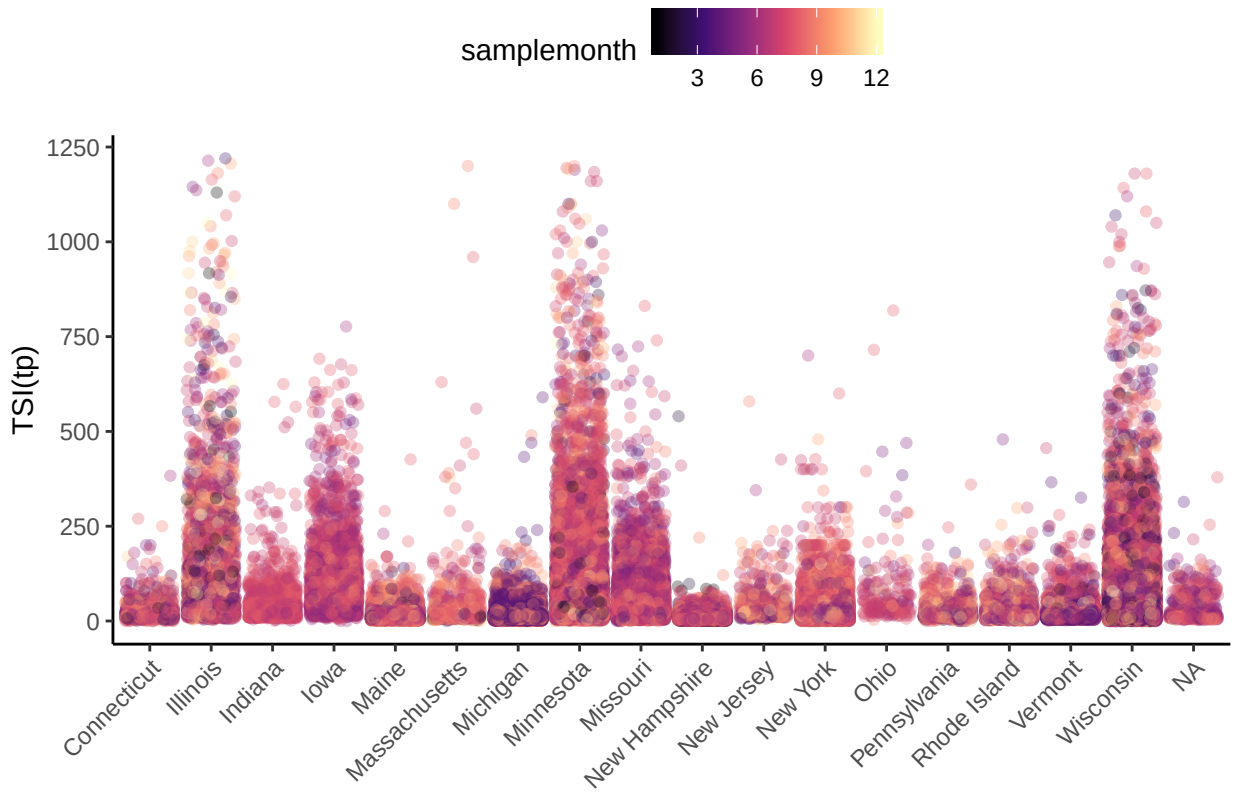
```
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1))
print(stateTNjitter)
```

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




```
stateTPjitter <- ggplot(LAGOSNandP, aes(x = state_name, y = tp, color = samplemonth)) +
 geom_jitter(alpha = 0.3) +
  labs(x = "", y = "TSI(tp)") +
  theme(legend.position = "top") +
  scale_color_viridis_c(option = "magma")+
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1))
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: Most: Iowa; Least: Vermont.

TP: Most: Illinois, Minnesota, Wisconsin; Least: New Hampshire

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

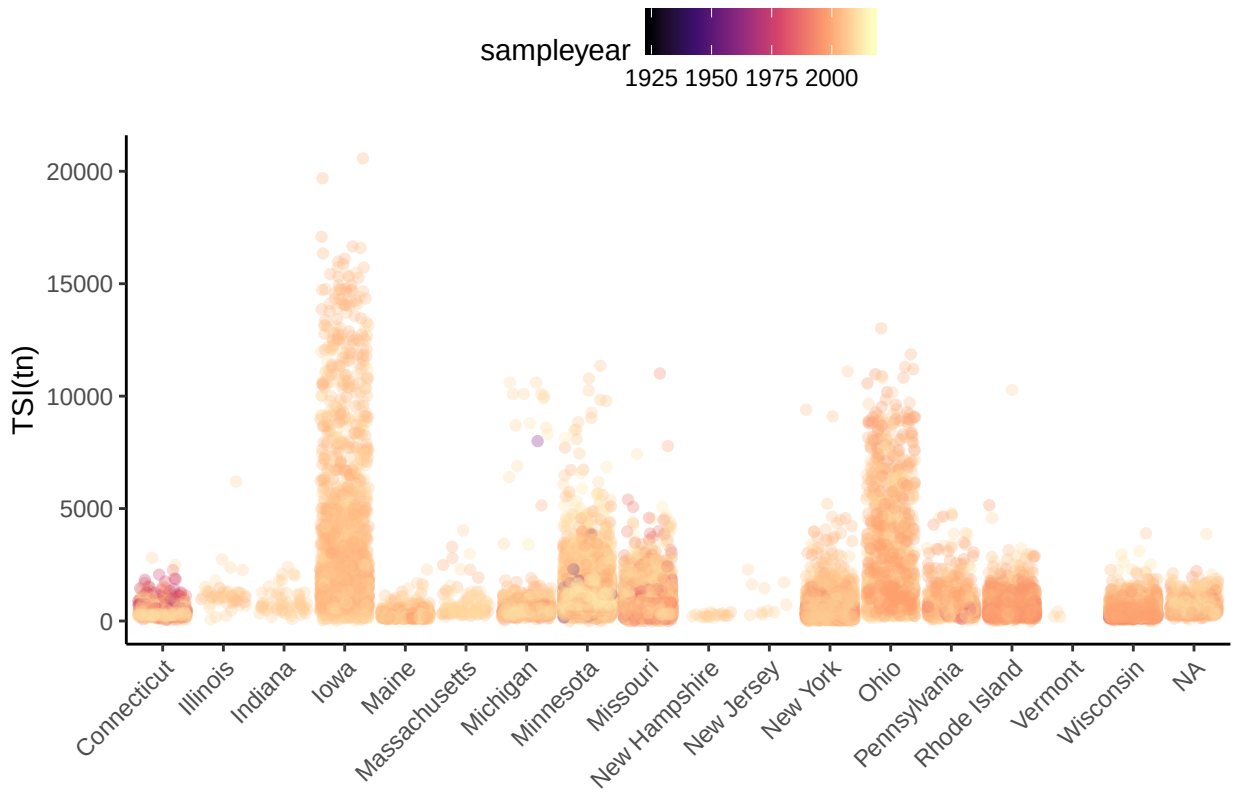
TN: It appears that the late summer months are the most extensively samples. There does seem to be some differing among states, as some seem to sample exclusively during the summer and some do so across multiple seasons.

TP: Again, it appears that the late summer months are the most extensively samples. There does seem to be some differing among states, as some seem to sample exclusively during the summer and some do so across multiple seasons.

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.

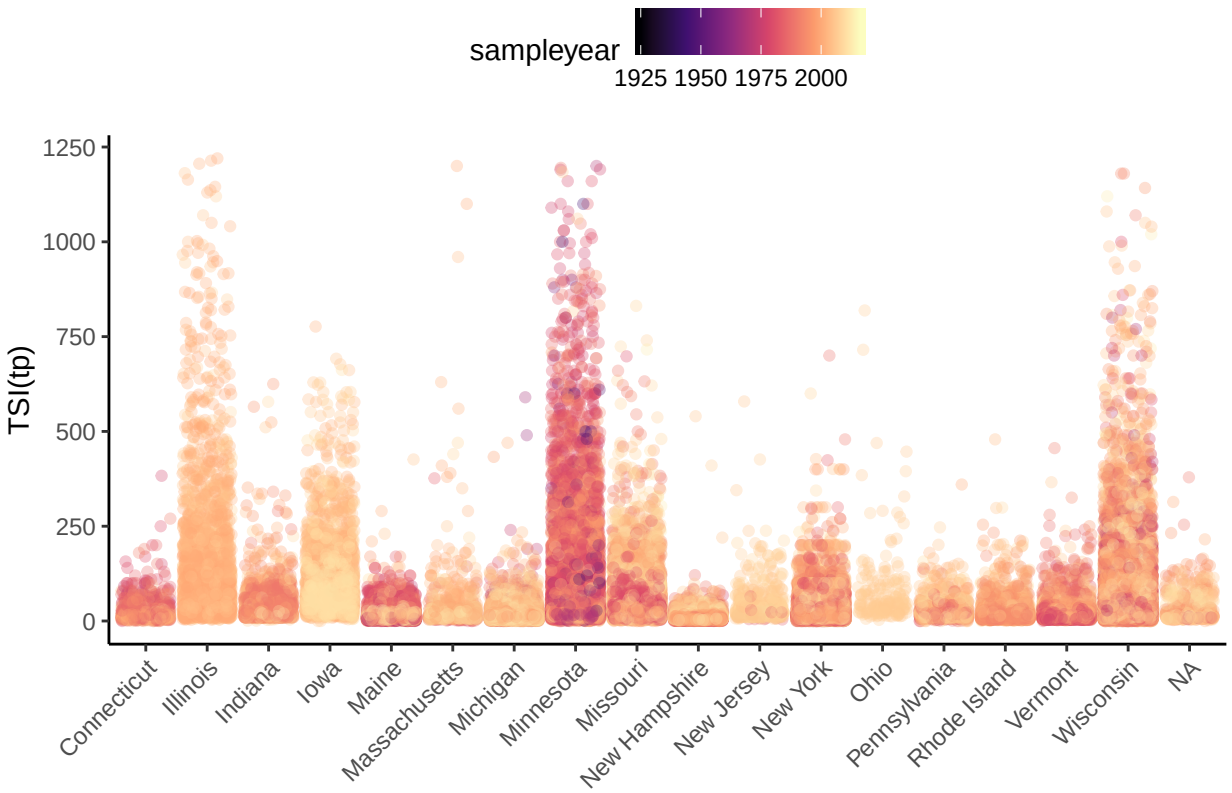
```
stateTNjitter.year <- ggplot(LAGOSNandP, aes(x = state_name, y = tn, color = sampleyear)) +
  geom_jitter(alpha = 0.3) +
  labs(x = "", y = "TSI(tn)") +
  theme(legend.position = "top") +
  scale_color_viridis_c(option = "magma") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1))
print(stateTNjitter.year)
```

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




```
stateTPjitter.year <- ggplot(LAGOSNandP, aes(x = state_name, y = tp, color = sampleyear)) +
 geom_jitter(alpha = 0.3) +
  labs(x = "", y = "TSI(tp)") +
  theme(legend.position = "top") +
  scale_color_viridis_c(option = "magma")+
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1))
print(stateTPjitter.year)
```

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




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

TN: It appears that the majority of samples are from 2000 to the present. It doesn't seem to differ drastically between states.

TP: It appears that most of the samples are from 2000 to the present, but there's a decent amount that seem to date back to the mid-70s. The TP samples have more more variation between states in regards to when samples were collected.

## Reflection

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

    Different variables can indicate very different levels of eutrophication. The time of year data is sampled can play a large role on the trends seen in data.

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

    The trophic data where we looked at the proportion of total observations which are considered eutrophic or hypereutrophic. (Count data statistics) States that sampled more frequently and sampled across seasons have larger ranges in their concentration data. (Jitter plots showed us this)

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

    Yes, having the opportunity to analyze real-world data allows for more critical thinking skills to be used to understnad trends and apply theoretical expectations to real-world data.

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

    There real-world data illustrates how messy things can be, and how sample size and timing (sampling decisions are made by humans!) can play a huge factor in the types of trends you see.