# Assignment 5: Water Quality in Lakes

*Rachel Bash*

## 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] "C:/Users/19524/Documents/DUKE/Hydrologic Data Analytics/Hydrologic_Data_Analysis/Assignments"
```

```
library(tidyverse)
library(lubridate)
library(LAGOSNE)

theme_set(theme_classic())
options(scipen = 100)

LAGOSdata <- lagosne_load()
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"
```

```
LAGOStrophic <- read.csv("../Data/Processed/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.

```
chla.count <- count(LAGOStrophic, trophic.class, name="Chla")
secchi.count <- count(LAGOStrophic, trophic.class.secchi, name="secchi")
tp.count <- count(LAGOStrophic, trophic.class.tp, name="tp")

class.list <- cbind(chla.count, secchi.count, tp.count) %>%
  select(trophic.class, Chla, secchi, tp)

colnames(class.list)[1] <- "class"

print(class.list)
```

```
##            class  Chla secchi    tp
## 1      Eutrophic 13317  12207 10158
## 2 Hypereutrophic  6392   2766  2658
## 3    Mesotrophic  4566   6804  6636
## 4   Oligotrophic  1560   4058  6383
```

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.proportion <- sum(class.list$Chla[3:4])/sum(class.list$Chla)
trophic.class.proportion
```

```
## [1] 0.2371202
```

```
trophic.class.sechhi.proportion <- sum(class.list$secchi[3:4])/sum(class.list$secchi)
trophic.class.sechhi.proportion
```

```
## [1] 0.4204374
```

```
trophic.class.tp.proportion <- sum(class.list$tp[3:4])/sum(class.list$tp)
trophic.class.tp.proportion
```

```
## [1] 0.5039288
```

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

> Secchi disk is the most conservative in its designation of eutrophic conditions. This is likely because this is done with human hands and eyes, and so collection and observation can vary largely on the person performing the task. It is probably the least precise out of the three TSI

indexes. We know it is the most conservative because it has the lowest proportion of points labeled either hypereutrophic or eutrophic, indicating that fewer samples would be considered those if the TSI index was based on secchi readings.

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.

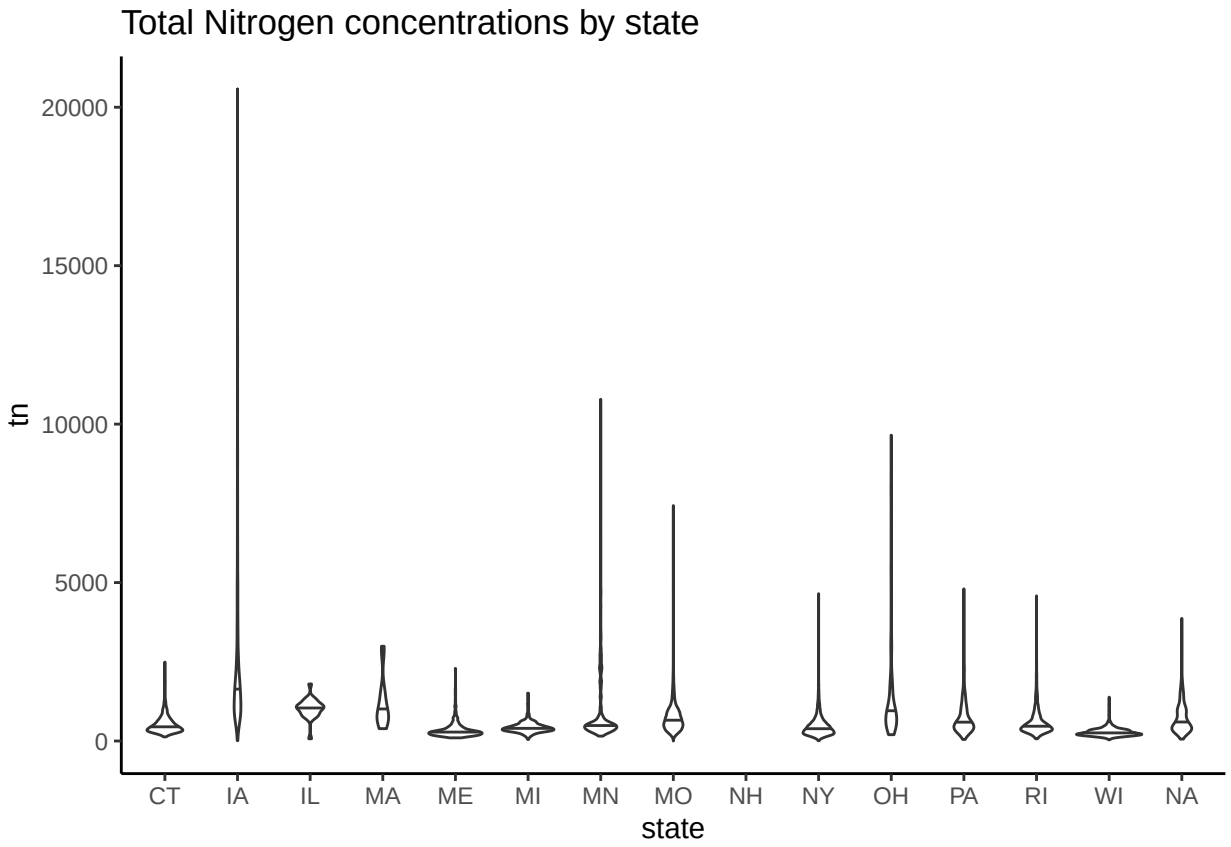
## Nutrient Concentrations

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.

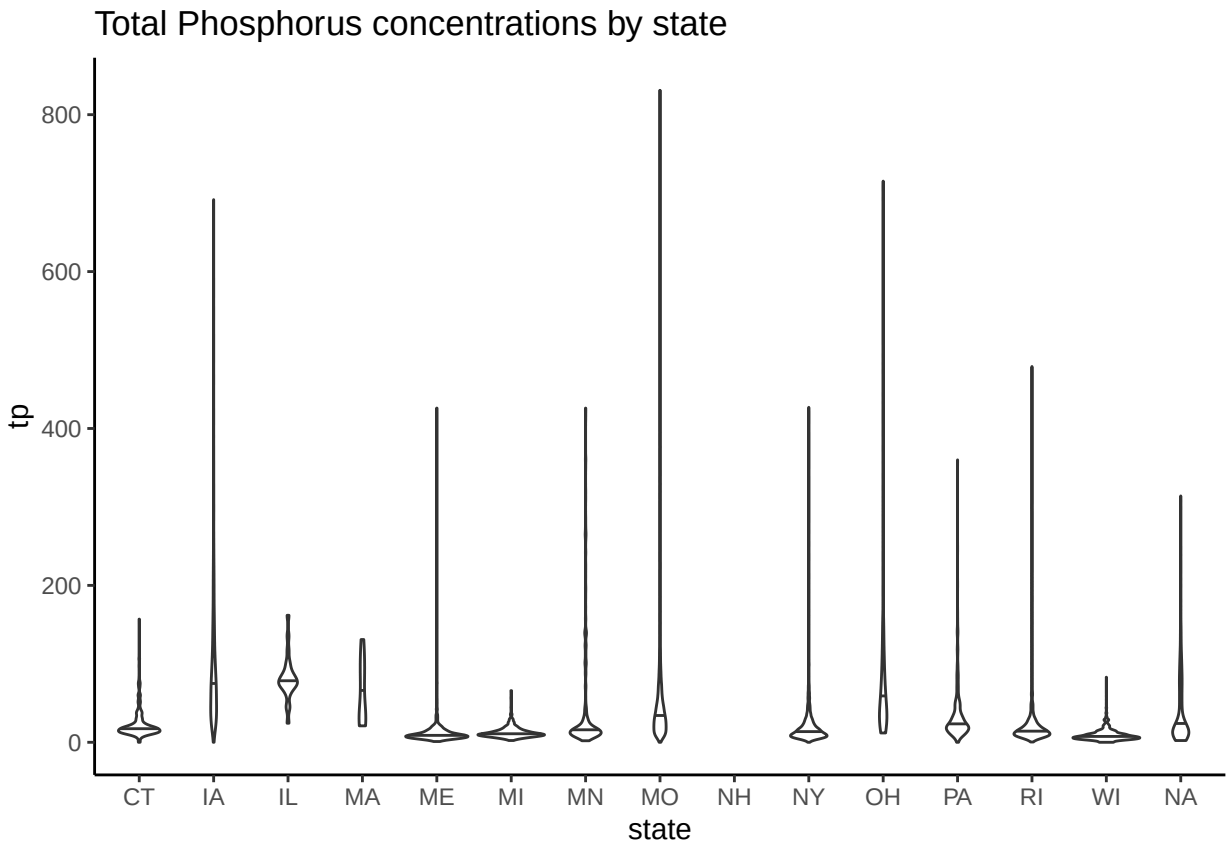
```
LAGOSNandP <- LAGOStrophic %>%
  select(lagoslakeid, sampledate, tn, tp, state, state_name) %>%
  mutate(sampleyear = year(sampledate)) %>%
  mutate(samplemonth = month(sampledate))
```

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

```
TNstate.violin <- ggplot(LAGOSNandP, aes(x=state, y=tn)) +
  geom_violin(draw_quantiles=0.5) +
  ggtitle("Total Nitrogen concentrations by state")
print(TNstate.violin)
```



```
TPstate.violin <- ggplot(LAGOSNandP, aes(x=state, y=tp)) +
  geom_violin(draw_quantiles=0.5)+
  ggtitle("Total Phosphorus concentrations by state")
print(TPstate.violin)
```




```
#tapply(LAGOSNandP$tn,LAGOSNandP$state,median)

#tapply(LAGOSNandP$tn, LAGOSNandP$state,summary)
```

Which states have the highest and lowest median concentrations?

TN: Iowa has the highest, Wisconsin has the lowest

TP: Illinois has the highest, Wisconsin has the lowest

Which states have the highest and lowest concentration ranges?

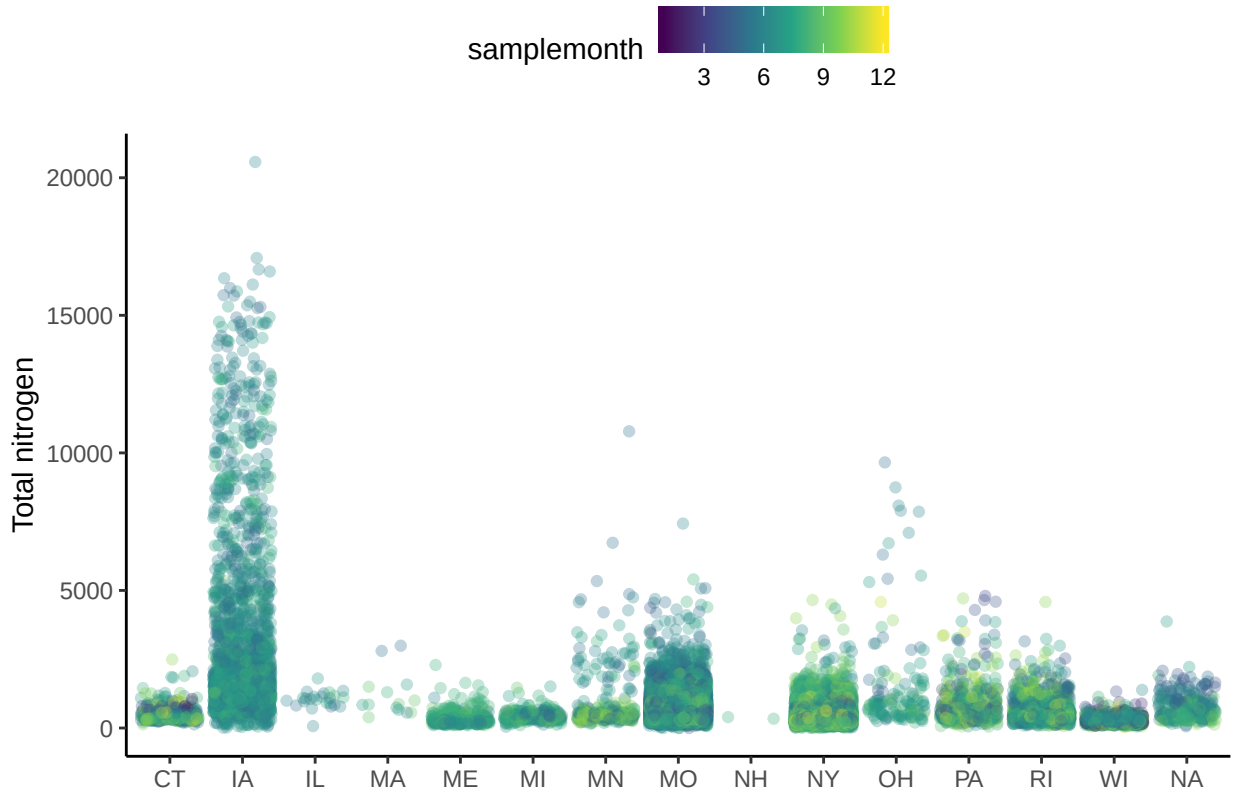
TN: Iowa has the biggest range, Wisconsin has the smallest range

TP: Missouri has the biggest range, Michigan has the smallest range

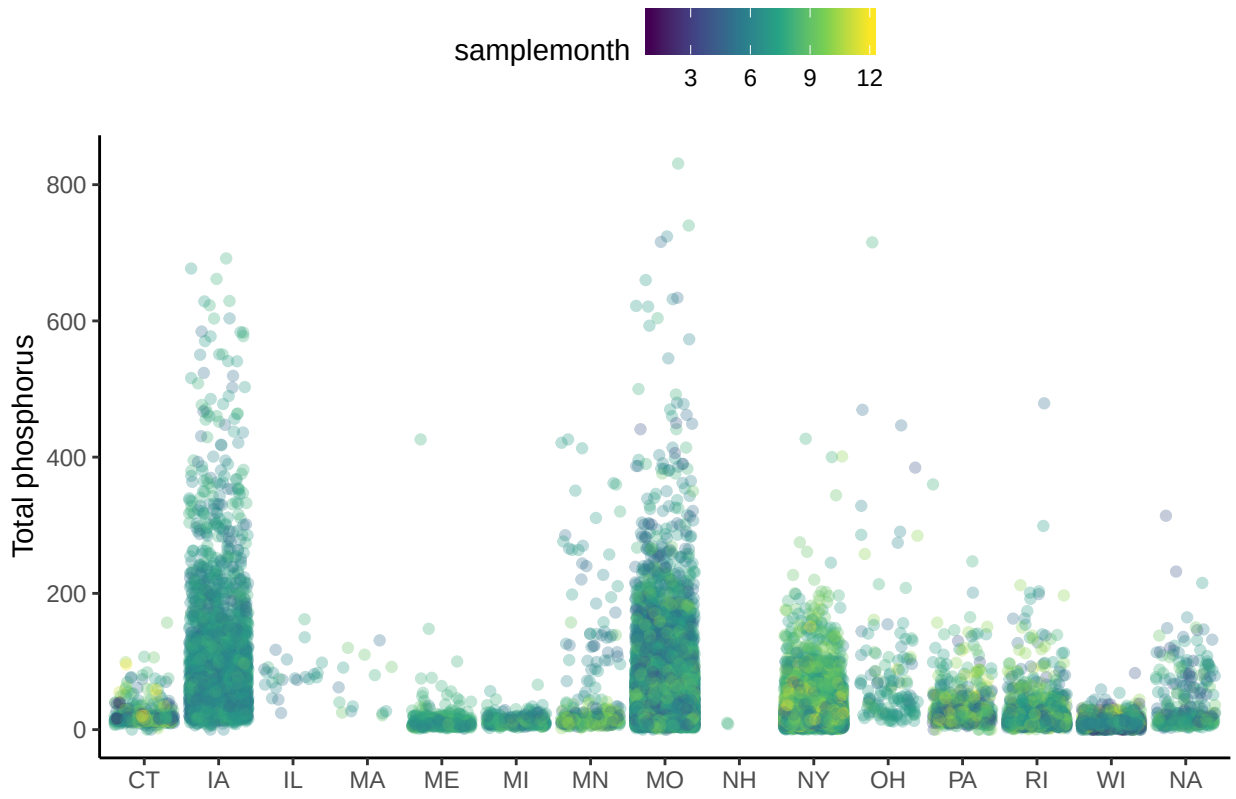
10. Create two jitter plots comparing TN and TP concentrations across states, with samplemonth as the color. Choose a color palette other than the ggplot default.

```
TNstate.jitter <- ggplot(LAGOSNandP, aes(x = state, y = tn,
                                         color = samplemonth)) +
  geom_jitter(alpha = 0.3) +
  labs(x = "", y = "Total nitrogen") +
  theme(legend.position = "top") +
```

```
  scale_color_viridis_c(option = "viridis")
print(TNstate.jitter)
```




```
TPstate.jitter <- ggplot(LAGOSNandP, aes(x = state, y = tp,
                                         color = samplemonth)) +
  geom_jitter(alpha = 0.3) +
  labs(x = "", y = "Total phosphorus") +
  theme(legend.position = "top") +
  scale_color_viridis_c(option = "viridis")
print(TPstate.jitter)
```




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

> TN: Iowa and Missouri have the most samples, as depicted by the number of points in the graph. Because Iowa seems to have the most amount of points, it is more likely that there will end up being greater variability among the possible points - making it most likely to have the biggest range.
>
> TP: Again, Iowa and Missouri have the most samples. Missouri has the largest range and the most amount of points, again showing that the more samples there are, the more tendency for greater variability in the data.

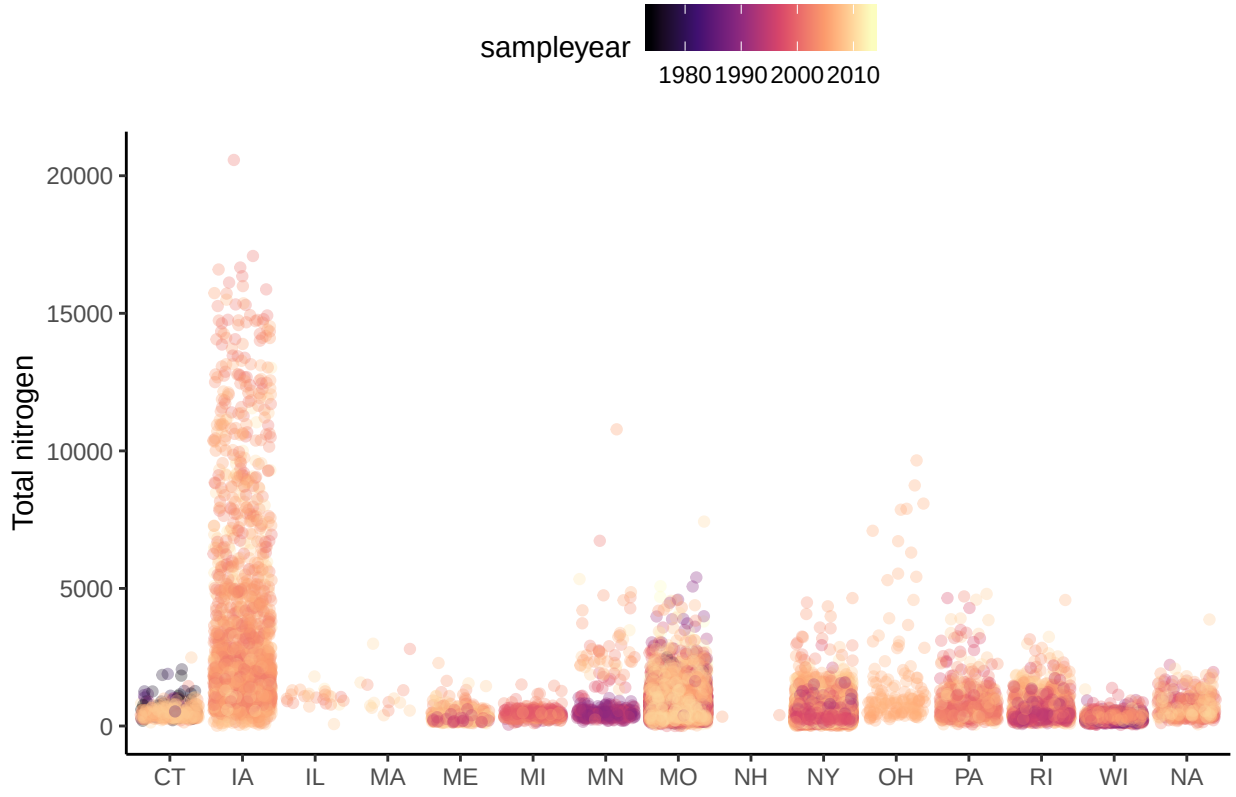
Which months are sampled most extensively? Does this differ among states?

> TN: The summer months are sampled the most extensively, with a little sampling in the early part and late part of each year. NY and PA have a lot of samples taken at the end of the year, while Missouri and Iowa have samples taken from the early part of the year and the summer.
>
> TP: The summer months are sampled the most extensively. A lot of the eastern states, like NY, RI, PA, and CT all have samples taken later in the year (yellow color). Iowa and Missouri have most of their points taken from April to June. It looks like CT and PA have the biggest range of sample months - ranging from the beginning of the year (dark blue) to the end (yellow).

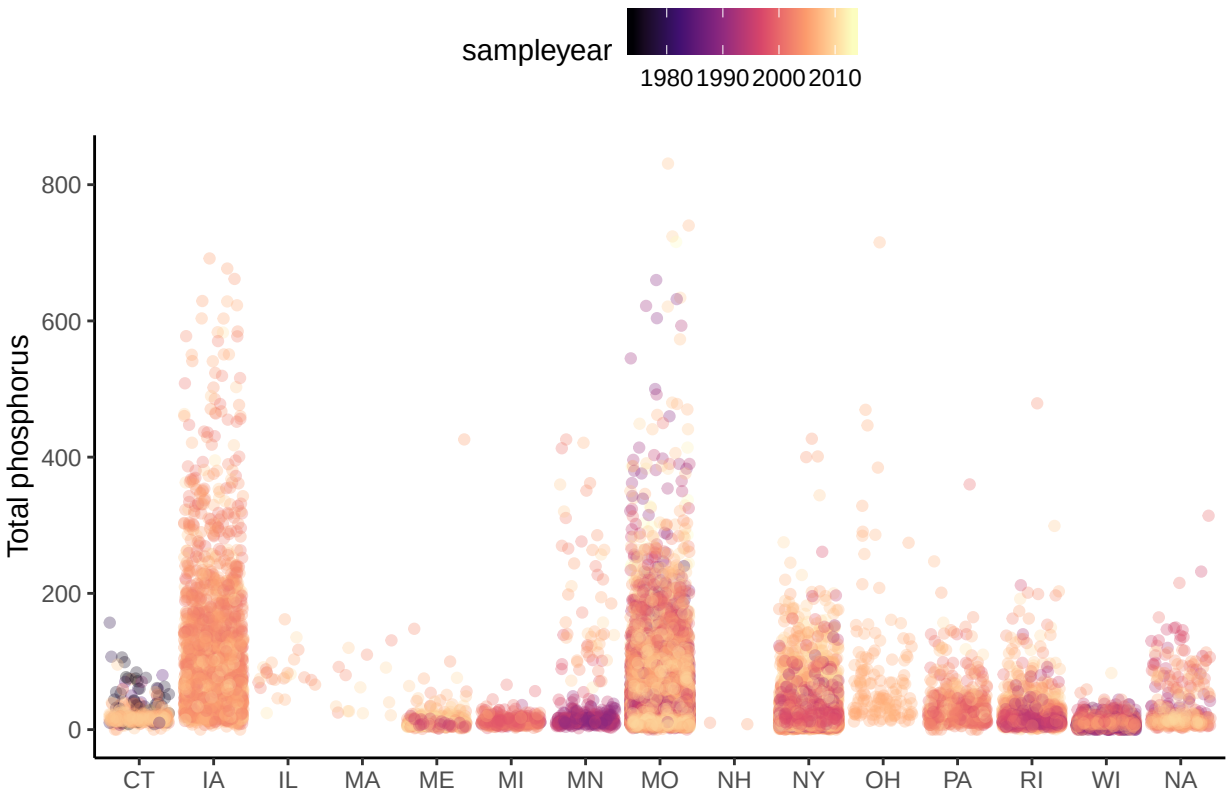
11. Create two jitter plots comparing TN and TP concentrations across states, with sampleyear as the color. Choose a color palette other than the ggplot default.

```
TNstate.jitter.year <- ggplot(LAGOSNandP, aes(x = state, y = tn,
                                              color = sampleyear)) +
  geom_jitter(alpha = 0.3) +
  labs(x = "", y = "Total nitrogen") +
  theme(legend.position = "top") +
```

```
  scale_color_viridis_c(option = "magma")
print(TNstate.jitter.year)
```




```
TPstate.jitter.year <- ggplot(LAGOSNandP, aes(x = state, y = tp,
                                              color = sampleyear)) +
  geom_jitter(alpha = 0.3) +
  labs(x = "", y = "Total phosphorus") +
  theme(legend.position = "top") +
  scale_color_viridis_c(option = "magma")
print(TPstate.jitter.year)
```




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

TN: Most samples were taken in the 2000s decade. MN and WI and CT have a fair amount of samples that were taken earlier, like in the 1980s. NY and MI and RI have samples taken in the 1990s. Iowa, with the most amount of samples, was almost exclusively taken in the 2000s.

TP: There is a bit more variation in this one, but again most of the samples were taken in the 2000s. MO and NY have a good mixture of many decades, while MN and MI have most in the 1980s and 1990s respectively. It is interesting to notice that MN's lower values were taken in the 1980s, but any P levels above about 50 were taken in the 2000s or later. You can also see a similar time trend in CT, where values above 50 were taken before about 1990, but anything below 50 was sampled in the 2000s or later.

## Reflection

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

    Year and time of year are important water quality measures that tell you a lot about the water system you are studying, especially as water quality measures change over time. You can glean all the information worth knowing from one plot; it is helpful to visualize the data in multiple ways to get the full picture. Lastly, a water quality parameter, such as trophic class, can change based on what you use to define it.

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

    It was helpful to differentiate year and season changes with the different jitter plots produced. They both told a different story than the violin plots did.

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

    I am not sure I learned anything too new about water quality this lesson relative to what a theory-based lesson would have taught me. This assignment was more about the coding than water quality knowledge.

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

    I didn't expect there to be such a discrepancy in the number of samples taken by state. There are so few samples taken in Illinois than there are in Missouri, for example. I have to be careful with how strongly to draw conclusions based on this data, because a lot of information for certain states could be missing, which may be hiding the real story of the data.