Assignment 5: Water Quality in Lakes

Gaby Garcia

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.

Set Working Directory

```
setwd("/Users/gabrielagarcia/Desktop/Hydrologic Data Analysis/Hydrologic_Data_Analysis/Data_Processed")
```

Load Packages

```
library(tidyverse)
library(lubridate)
library(LAGOSNE)
library(corrplot)
library(cowplot)
library(dplyr)
library(viridis)
```

###Set GGPlot Theme

```
gabytheme <- theme_bw(base_size = 22) +
  theme(plot.title=element_text(face="bold", size="29", color="IndianRed3", hjust=0.5),
        axis.title=element_text(size=22, color="black"),
axis.text = element_text(face="bold", size=18, color = "black"),
panel.background=element_rect(fill="white", color="darkblue"),
panel.border = element_rect(color = "black", size = 2),
legend.position = "top", legend.background = element_rect(fill="white", color="black"),
            legend.key = element_rect(fill="transparent", color="NA"))

theme_set(gabytheme)
```

Remove Scientific Notation from Data Frame

```
options(scipen = 100) ###To remove scientific notation from data frame columns
```

Load LAGOSdata database

```
library(LAGOSNE)
setwd("/Users/gabrielagarcia/Desktop/Hydrologic Data Analysis/Hydrologic_Data_Analysis/Data_Processed")
LAGOStrophic<- read.csv("LAGOStrophic.csv")
#lagosne_get(dest_folder = LAGOSNE:::lagos_path(), overwrite = TRUE)
LAGOSdata<-lagosne_load()
```

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).

Add Trophic Indices columns for Secchi and Phosphorus

```
LAGOStrophic<-LAGOStrophic %>%
  mutate(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"))))
```

Make Trophic Class Columns a Factor

```
LAGOStrophic$trophic.class.secchi<-as.factor(LAGOStrophic$trophic.class.secchi)
LAGOStrophic$trophic.class.tp<-as.factor(LAGOStrophic$trophic.class.tp)
```

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.

Create Trophic Count Tables

```
Chltrophiccount<-count(LAGOStrophic, trophic.class)
Secchitrophiccount<-count(LAGOStrophic, trophic.class.secchi)
tptrophiccount<-count(LAGOStrophic, trophic.class.tp)
```

Print Preliminary Tables

```
Chltrophiccount
```

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

```
## 4 Oligotrophic    3298
```

```
Secchitrophiccount
```

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

```
tptrophiccount
```

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

Create Trophic Count Table

```
TrophicCount<-cbind(Chltrophiccount, Secchitrophiccount$n, tptrophiccount$n)%>%
  rename(chlorophyll=n, secchi="Secchitrophiccount$n", tp="tptrophiccount$n")

print(TrophicCount)
```

```
##    trophic.class chlorophyll secchi    tp
## 1      Eutrophic       41861  28659 24839
## 2 Hypereutrophic       14379   5099  7228
## 3    Mesotrophic       15413  25083 23023
## 4   Oligotrophic        3298  16110 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 Table

```
Proportion<-TrophicCount%>%mutate(chlorophyll=round(chlorophyll/sum(chlorophyll), 3),
                                  secchi=round(secchi/sum(secchi), 3),
                                  tp=round(tp/sum(tp), 3))
Proportion
```

```
##    trophic.class chlorophyll secchi    tp
## 1      Eutrophic       0.559  0.382 0.331
## 2 Hypereutrophic       0.192  0.068 0.096
## 3    Mesotrophic       0.206  0.335 0.307
## 4   Oligotrophic       0.044  0.215 0.265
```

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.

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

> Total phosphorus is the most conservative parameter in its designation of eutrophic conditions. Phosphorus is the primary limiting factor for plant growth in several freshwater ecosystems. According to Ullman's Encyclopedia of Industrial Chemistry, the availability of phosphorus promotes excessive plant growth and decay, while favoring algae and plankton over other plants, thus reducing overall water quality.

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.

Obtain Metadata, State, and Nutrient Information

```
LAGOSnutrient <- LAGOSdata$epi_nutr


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

Tell R to treat lakeid as a factor, not a numeric value

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

Join locus and state data frames

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

Create LAGOStrophic data frame from LAGOSnutrient and LAGOSlocations dataframes

```
LAGOStrophicfinal<-
  left_join(LAGOSnutrient, LAGOSlocations, by = "lagoslakeid") %>%

  select(lagoslakeid, sampledate, tp, tn, state, state_name) %>%
  mutate(sampleyear = year(sampledate),
         samplemonth = month(sampledate),

         season = as.factor(quarter(sampledate, fiscal_start = 12)))
```

Create LAGOStrophicfinal dataframe

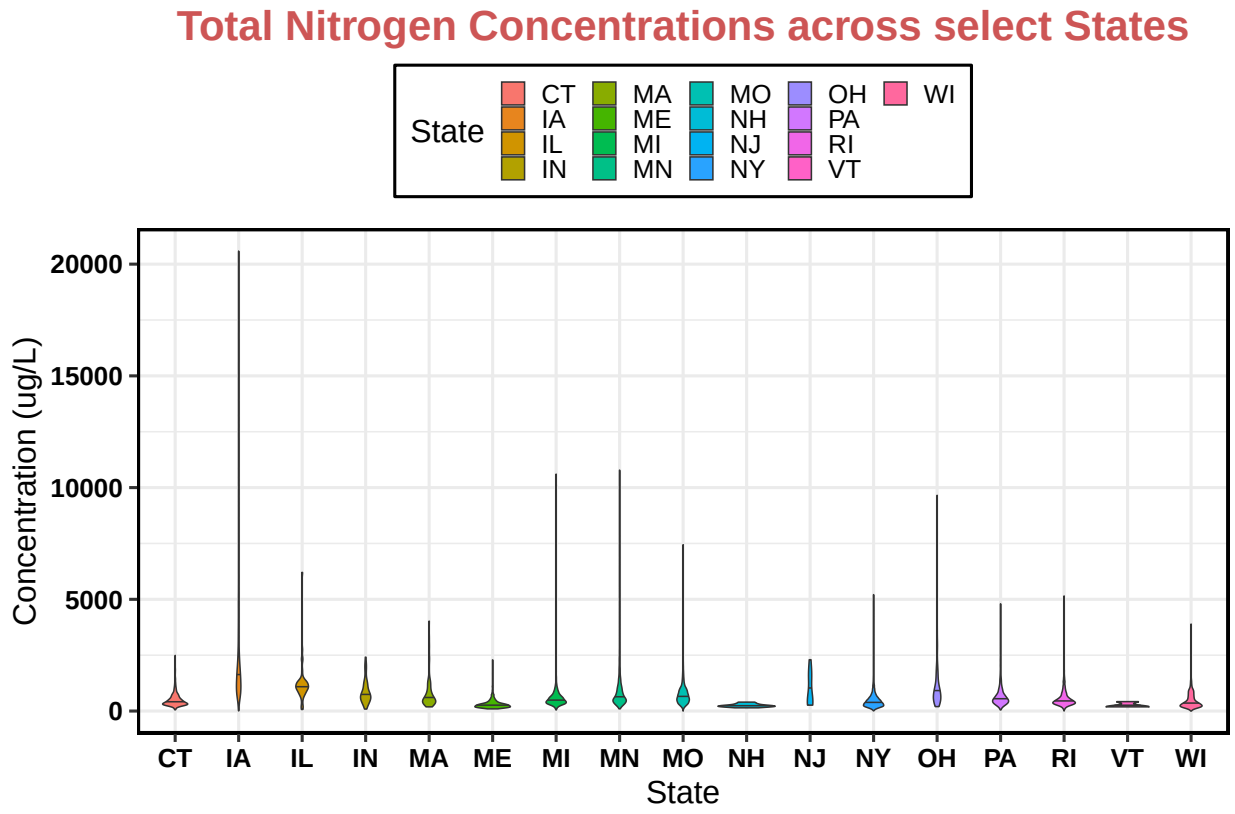
```
library(tidyr)
LAGOSNandP<-LAGOStrophicfinal%>%
  select(lagoslakeid, sampledate, tn, tp, state, state_name)%>%
  mutate(sampleyear = year(sampledate),
         samplemonth = month(sampledate))%>%
  drop_na(tn:state)
```

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

TN Concentrations

```
TNViolin<-ggplot(LAGOSNandP,
                          aes(x =factor(state), y =tn)) +
  geom_violin(trim=TRUE, draw_quantiles = c(0.5),
              aes(fill=factor(state))) +
  labs(x="State", y="Concentration (ug/L)",
       title="Total Nitrogen Concentrations across select States")+
  guides(fill=guide_legend(title="State"))+
  gabytheme

print(TNViolin)
```

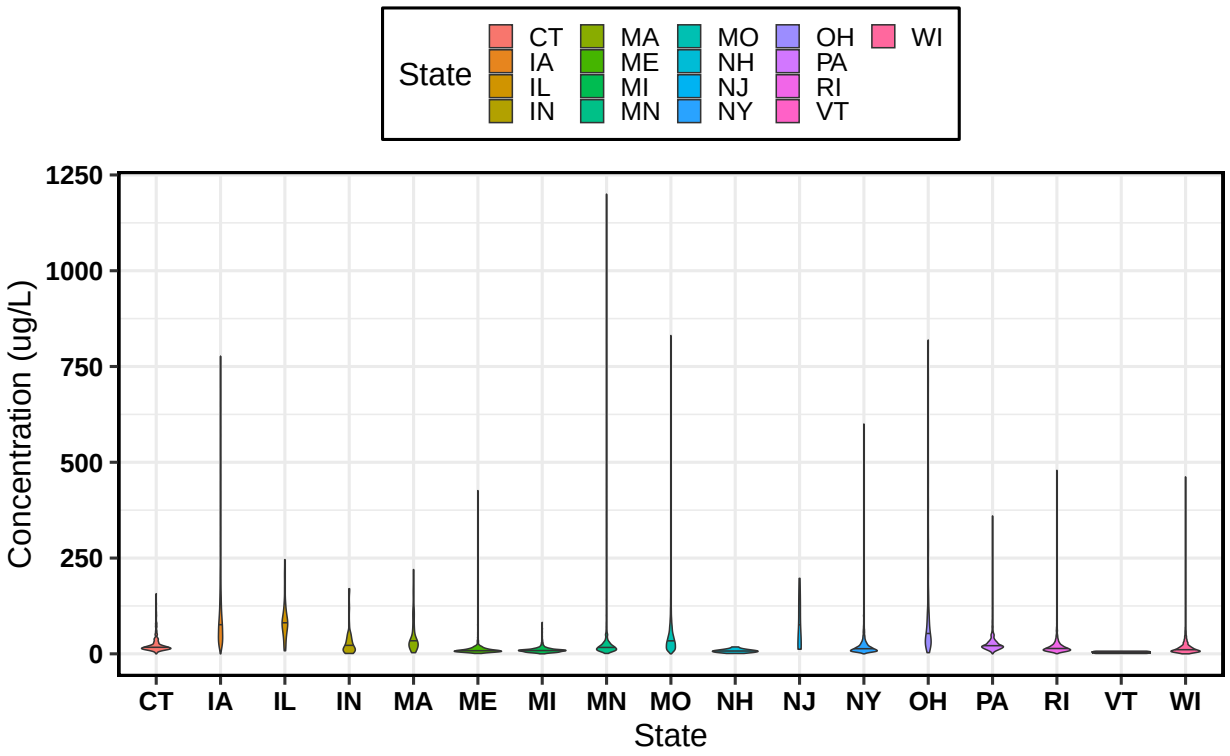


TP Concentrations

```
TPViolin<-ggplot(LAGOSNandP,
                          aes(x =factor(state), y =tp)) +
  geom_violin(trim=TRUE, draw_quantiles = c(0.5),
              aes(fill=factor(state))) +
  labs(title="Total Phosphorus Concentrations across select States",
       x="State",
       y="Concentration (ug/L)")+
```

```
  guides(fill=guide_legend(title="State"))+
  gabytheme

print(TPViolin)
```



LAGOS Summary

```
LAGOSSummary<-LAGOSNandP%>%
  group_by(state)%>%
  summarize(mediantn=median(tn, na.rm=T),
            mediantp=median(tp, na.rm=T),
            rangetn=max(tn, na.rm=T)-min(tn, na.rm=T),
            rangetp=max(tp, na.rm=T)-min(tp, na.rm=T))

LAGOSSummary
```

```
## # A tibble: 17 x 5
##    state mediantn mediantp rangetn rangetp
##    <chr>    <dbl>    <dbl>   <dbl>   <dbl>
##  1 CT         410       17     2435     157.
##  2 IA        1628.      75.9  20564.    776.
##  3 IL        1084.      79.3   6133     238
##  4 IN         714       17     2323     169
##  5 MA         532       30     3834     217
##  6 ME         261        8.3   2193     425
##  7 MI         493        8.7  10552.     82
```

```
##  8 MN        622      17    10683    1199
##  9 MO        660      34     7430     831
## 10 NH        234       8      255      17
## 11 NJ        585      31     2033     185
## 12 NY        390      14     5200     600
## 13 OH        854.     46.6   9454     816
## 14 PA        550      21     4750     360
## 15 RI        455      14     5146     479.
## 16 VT        204       4      234       6
## 17 WI        351      11     3893     462
```

Which states have the highest and lowest median concentrations?

TN: Iowa has the highest median concentration while Wisconsin has the lowest median concentration.

TP: Illinois has the highest median concentration while Wisconsin has the lowest median concentration.

Which states have the highest and lowest concentration ranges?

TN: Iowa has the highest nitrogen concentration range, while Vermont has the lowest range.

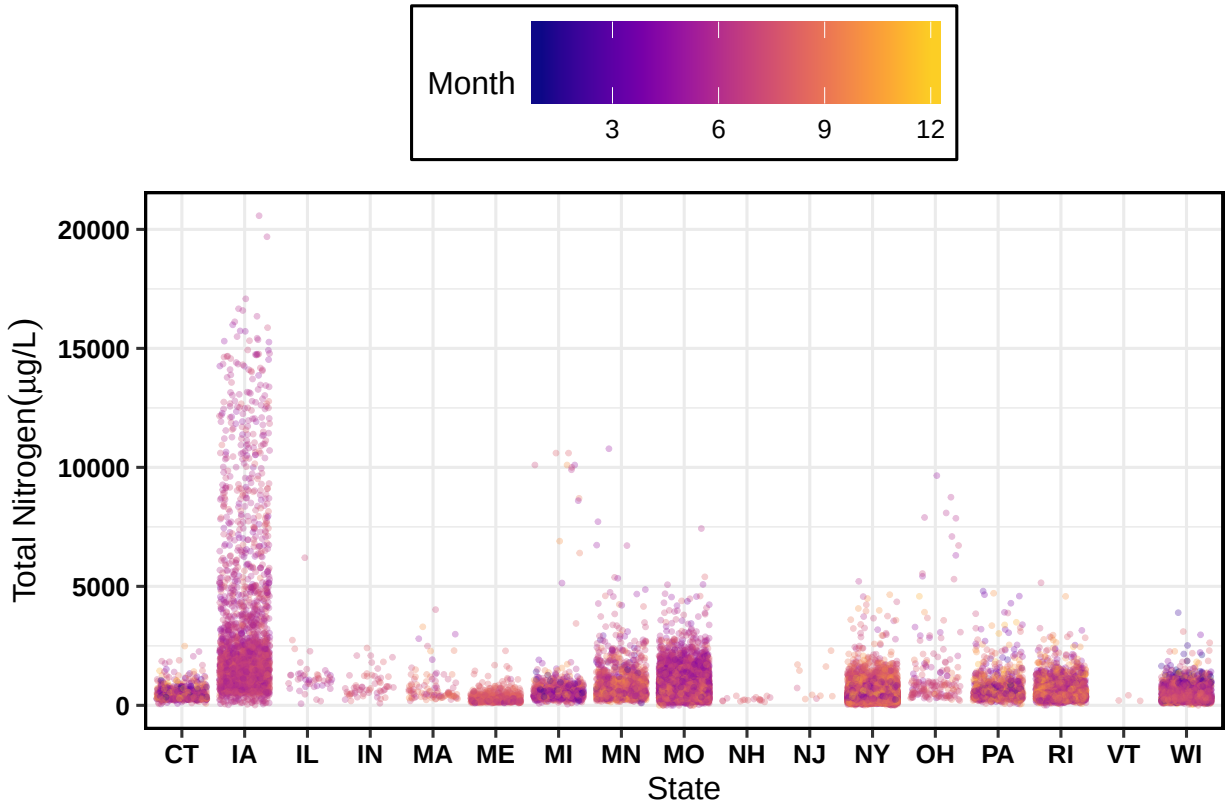
TP: Minnesota has the highest phosphorus concentration range, while Vermont has the lowest range.

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.

Total Nitrogen Jitter Plot Across States by Sample Month

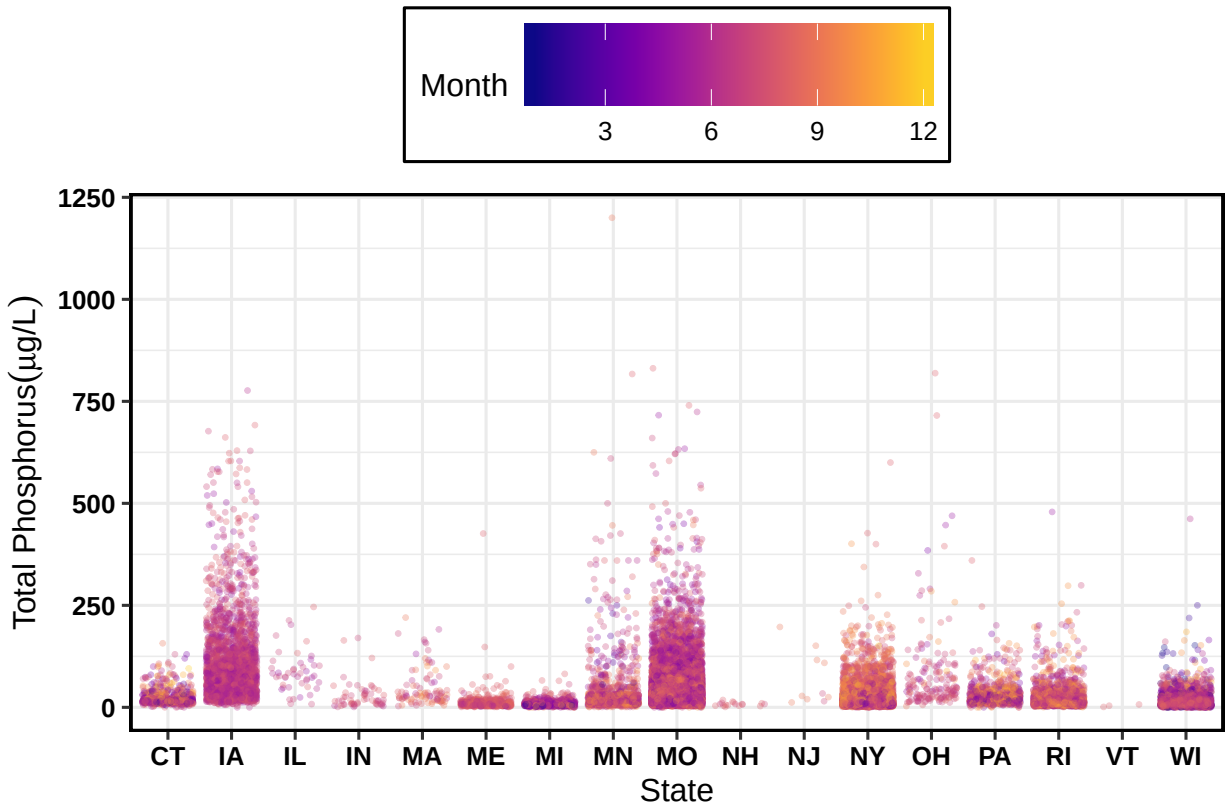
```
TNJitterPlot<-ggplot(LAGOSNandP, aes(x=state, y=tn,
                                     color=samplemonth))+
  geom_jitter(alpha=0.3)+
  scale_color_viridis(option="plasma", end=0.9)+
  labs(x="State", y=expression("Total Nitrogen"(mu*"g/L")),
       color="Month")+
  theme(legend.key.size = unit(2, "cm"))


print(TNJitterPlot)
```



Total Phosphorus Jitter Plot across States by Sample Month

```
TPJitterPlot<-ggplot(LAGOSNandP, aes(x=state, y=tp,
                                     color=samplemonth))+
  geom_jitter(alpha=0.3)+
  scale_color_viridis(option="plasma", end=0.9)+
  labs(x="State", y=expression("Total Phosphorus"(mu*"g/L")),
       color="Month")+
   theme(legend.key.size = unit(2, "cm"))
print(TPJitterPlot)
```



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

Nitrogen Count

```
TNState<-LAGOSNandP%>%
  select(state, tn)%>%
  count(state)
TNState
```

```
## # A tibble: 17 x 2
##    state     n
##    <chr> <int>
##  1 CT      636
##  2 IA     2638
##  3 IL       46
##  4 IN       57
##  5 MA       93
##  6 ME      633
##  7 MI      877
##  8 MN     1341
##  9 MO    11412
## 10 NH       17
## 11 NJ       10
## 12 NY     7715
## 13 OH      166
## 14 PA      983
## 15 RI     2753
```

```
## 16 VT        3
## 17 WI     2336
```

Phosphorus Count

```
TPState<-LAGOSNandP%>%
  select(state, tp)%>%
  count(state)
TPState
```

```
## # A tibble: 17 x 2
##    state     n
##    <chr> <int>
##  1 CT      636
##  2 IA     2638
##  3 IL       46
##  4 IN       57
##  5 MA       93
##  6 ME      633
##  7 MI      877
##  8 MN     1341
##  9 MO    11412
## 10 NH       17
## 11 NJ       10
## 12 NY     7715
## 13 OH      166
## 14 PA      983
## 15 RI     2753
## 16 VT        3
## 17 WI     2336
```

TN: Missouri, New York, Rhode Island, and Iowa seem to have the most sample points. Iowa's sample points have the highest range in nitrogen concentrations, with some samples exceeding 15,000 ug/L. This could have impacted Iowa's total nitrogen ranges from #9 because Iowa is ranked second in the United States for agricultural production, especially of corn. Therefore, I would expect the heavy use of fertilizers during the growing season, leading to nutrient runoff.

TP: Missouri, New York, Rhode Island, and Iowa also have the most phosphorus sample points. Even though Missouri and New York have more phosphorus sample points, Iowa's sample points features a greater range in phosphorus concentrations.

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

TN: The summer months of May, June, July, and August were sampled more for nitrogen than the winter months. I know this because there are a far greater number of hot pink points corresponding with the summer color range. This is substantiated by the fact that typical agricultural growing seasons peak in mid-summer, with any required fertilizers already applied. Some states like Minnesota, Maine, and New York appear to have a sizeable proportion of sample points collected late in the year, while states such as Michigan and Missouri have a sizable proportion of sample points collected early in the year (based on my eyeballing the plot and seeing dark blue to purple points).

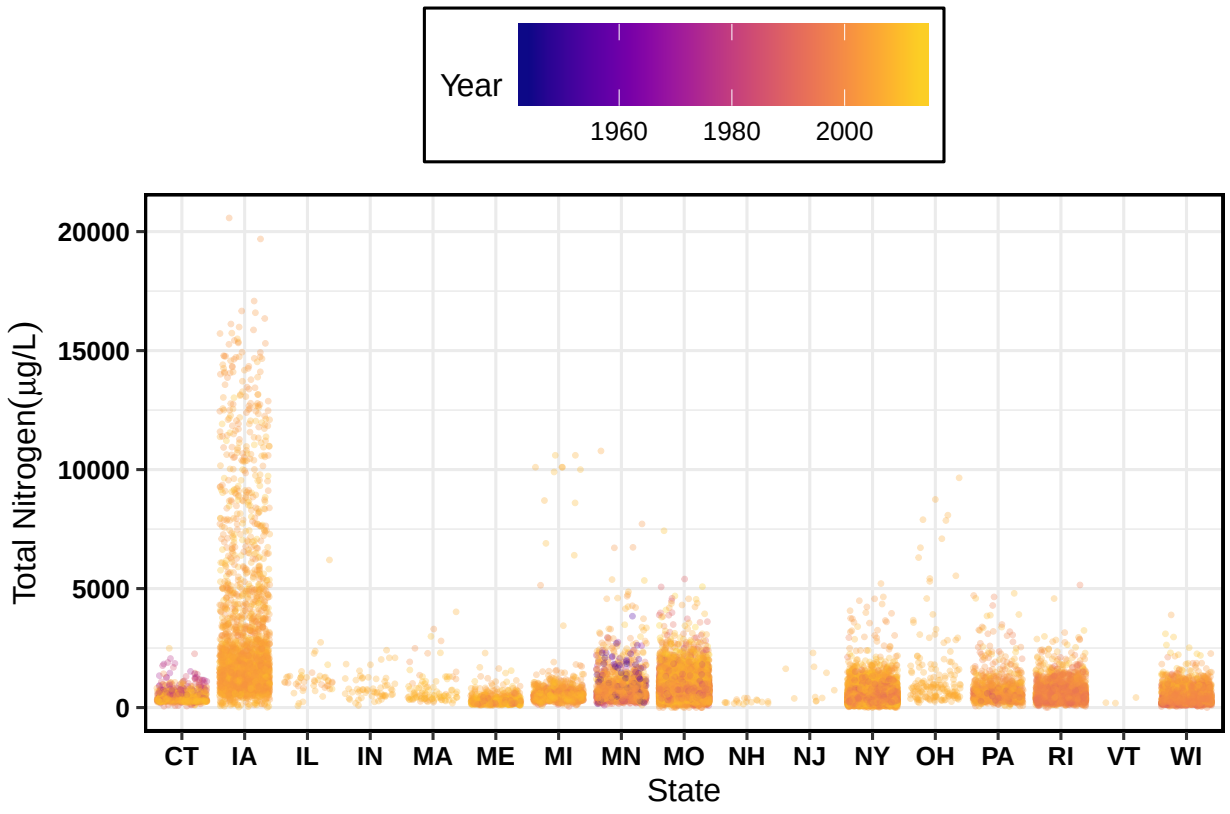
TP: As with total nitrogen, the summer months of May, June, July and August are sampled more for phosphorus than the winter months. I noticed that Indiana only sampled for Phosphorus in July and August, while again Michigan, Vermont, and Missouri have a sizeable proportion of points collected early in the year.

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.

Total Nitrogen Jitter Plot Across state by Sample Year

```
library(ggplot2)
library(viridis)
TNJitterPlotbyyear<-ggplot(LAGOSNandP,
                           aes(x=state, y=tn, color=sampleyear))+
  geom_jitter(alpha=0.3)+
  scale_color_viridis(option="plasma", end=0.9)+
  labs(x="State", y=expression("Total Nitrogen"(mu*"g/L")),
       color="Year")+
 theme(legend.key.size = unit(2, "cm"))

print(TNJitterPlotbyyear)
```

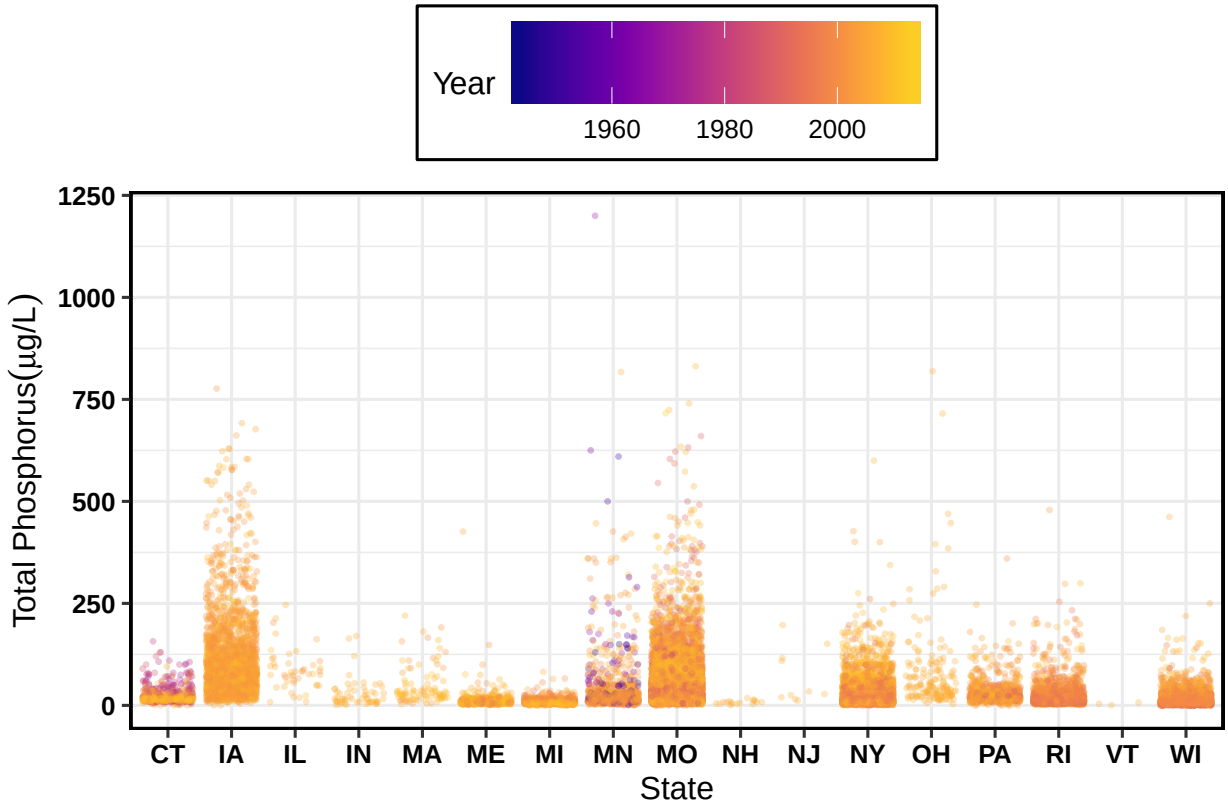


Total Phosphorus Jitter Plot Across state by Sample Year

```
library(ggplot2)
library(viridis)
TPJitterPlotbyyear<-ggplot(LAGOSNandP,
                           aes(x=state, y=tp, color=sampleyear))+
  geom_jitter(alpha=0.3)+
  scale_color_viridis(option="plasma", end=0.9)+
```

```
  labs(x="State", y=expression("Total Phosphorus"(mu*"g/L")),
       color="Year")+
 theme(legend.key.size = unit(2, "cm"))

print(TPJitterPlotbyyear)
```



Which years are sampled most extensively? Does this differ among states? ’ ### Sample Year Count

```
TNsampleyear<-LAGOSNandP%>%
  select(state, sampleyear)%>%
  count(sampleyear)
TNsampleyear
```

```
## # A tibble: 62 x 2
##    sampleyear     n
##         <dbl> <int>
##  1       1944     1
##  2       1945     2
##  3       1946     3
##  4       1947     2
##  5       1948     2
##  6       1949     5
##  7       1950     1
##  8       1953     8
##  9       1954     6
## 10       1955    14
## # ... with 52 more rows
```

TN: For all of the states, the years sampled most extensively were the 2000s and beyond, likely due to stricter nutrient pollution regulations in the 21st century. Minnesota and Connecticut visibly have a sizeable proportion of samples collected before the 1980s based the amount of visible points that range from blue to purple. Iowa has the highest number of sample points-the majority of them were taken in the 2000s and therefore has the highest range in concentrations. It is also interesting to note that while jitter plots may not fully represent the data due to hidden points, the sample points taken before the 1980s in Minnesota and Connecticut do not exceed 5000 ug/L.

TP: For total phosphorus samples, the states feature slightly more variability in their sample year than the total nitrogen samples, but the majority of the samples were collected in the 2000s and beyond. Minnesota has a higher proportion of samples visibly sampled before 1980 than the other states due to the visible points that range from blue to purple. Connecticut has less sample points than Minnesota but also a noteworthy proportion of points sampled before 1980.

Reflection

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

    Trophic indices are used to measure the amount of total biomass in an aquatic body, but they can be assessed by different metrics and the distinction between different trophic states is not necessarily clear. Based on sample analysis, the concentrations of nitrogen and phosphorus vary throughout time depending on the sampling state, and certain states clearly have stricter nutrient pollution standards (ex. regular aquatic monitoring) than others.

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

    Questions 6 and 7 showed how different trophic indices designate eutrophic conditions differently. The 4 jitter plots (2 modeling nitrogen and phosphorus over the sample months, 2 modeling nitrogen and phosphorus over the smple years) showed how long each individual state has been monitoring for nutrients to determine the lake's trophic state. The jitter plots also showed the changes in nutrients concentrations over time (provided data before 1980 was available).

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

    Yes, again being able to visualize the range and spread of water quality sampling points for different states helps me understand trends better because I am a visual learner. Also, it was interesting comparing the different types of graphs for the same manipulated dataset and processing and summarizing the different information conveyed by them.

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

    The real world data aligned with my expectations. States with high agricultural production in the Midwestern breadbasket region (Iowa, Missouri, Illinois) have more nutrient runoff and have a higher number of nutrient samples and trophic levels than states with presumbly lower agricultural production (New Hampshire, New Jersey). States with more sampling points are more likely to have a greater range of concentrations as well.