Assignment 5: Water Quality in Lakes

Yutao Gong

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/gongy/Documents/Hydrologic_Data_Analysis/Assignments"
```

```
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_bw())
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"))))

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.

```
LAGOStrophic %>%
        group_by(trophic.class) %>%
        summarise(Number_of_Observations = n()) %>%
        mutate(Proportion_of_TotalObservations = Number_of_Observations / sum(Number_o
f_Observations))
```

trophic.class <fctr>	Number_of_Observations <int>	Proportion_of_TotalObservations <dbl>
Eutrophic	41861	0.55851156
Hypereutrophic	14379	0.19184534
Mesotrophic	15413	0.20564102
Oligotrophic	3298	0.04400208

4 rows

```
LAGOStrophic %>%
        group_by(trophic.class.secchi) %>%
        summarise(Number_of_Observations = n()) %>%
        mutate(Proportion_of_TotalObservations = Number_of_Observations / sum(Number_o
f_Observations))
```

trophic.class.secchi <fctr>	Number_of_Observations <int>	Proportion_of_TotalObservatio <d
Oligotrophic	16110	0.21494(
Mesotrophic	25083	0.33465{
Eutrophic	28659	0.38236(
Hypereutrophic	5099	0.06803

4 rows

```
LAGOStrophic %>%
        group_by(trophic.class.tp) %>%
        summarise(Number_of_Observations = n()) %>%
        mutate(Proportion_of_TotalObservations = Number_of_Observations / sum(Number_o
f_Observations))
```

trophic.class.tp <fctr>	Number_of_Observations <int>	Proportion_of_TotalObservations <dbl>
Oligotrophic	19861	0.26498646

trophic.class.tp <fctr>	Number_of_Observations <int>	Proportion_of_TotalObservations <dbl>
Mesotrophic	23023	0.30717402
Eutrophic	24839	0.33140318
Hypereutrophic	7228	0.09643634

4 rows

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)?

```
# The proportion is shown as the third column "Proportion of Total Observations" in th
e last section
```

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

> The tp metric is the most conservative in designation of eutrophic conditions. It might be because we assume phosphorus is the limiting nutrient for phytoplankton growth (especially in summer times), which therefore constraints the potential eutrophic level given a certain phosphorus level.

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.

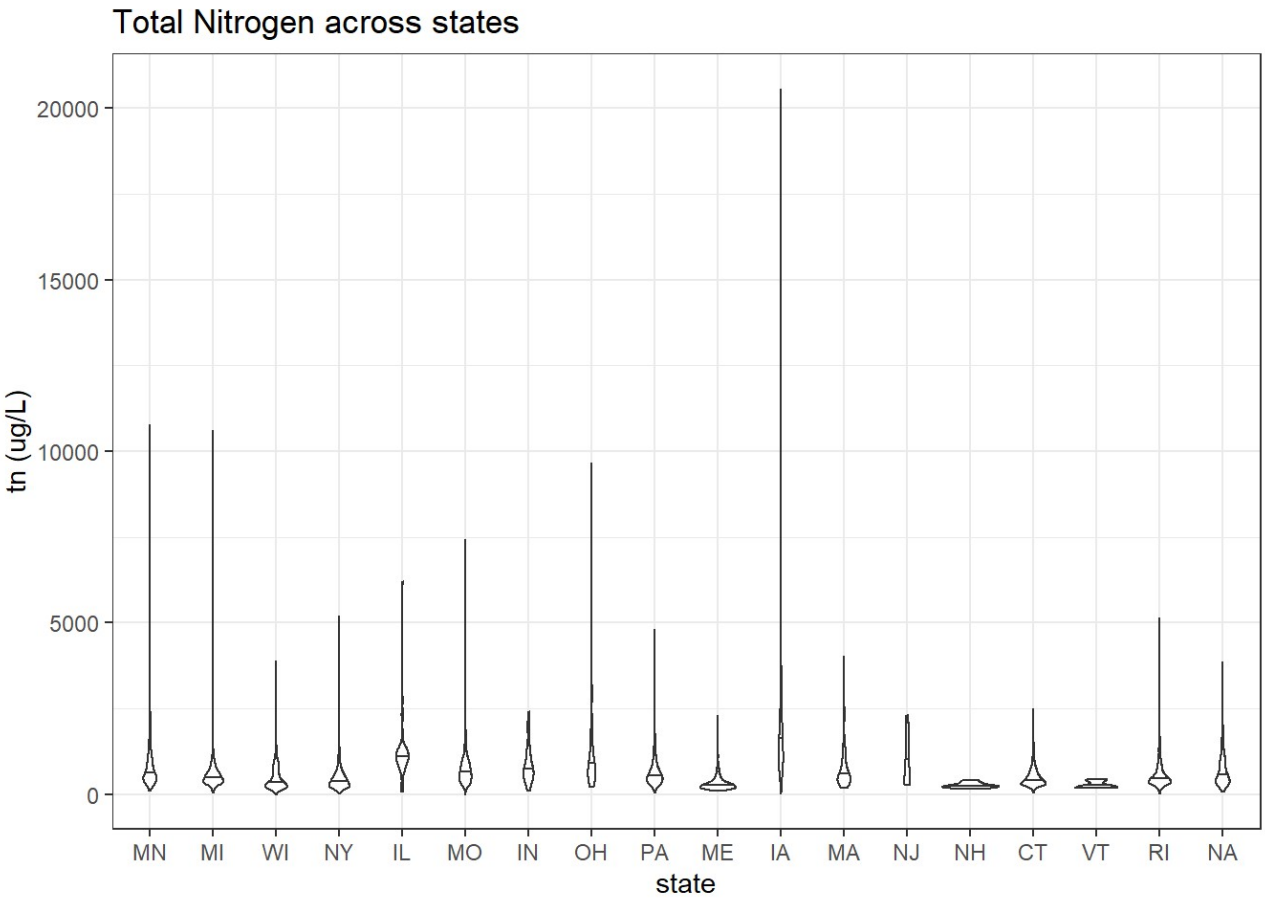
```
load(file = "../Data/Raw/LAGOSdata.rda")

# Exploring the data types that are available
LAGOSlocus <- LAGOSdata$locus # location
LAGOSstate <- LAGOSdata$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")
# by: column in common

# Order by number of lakes
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)) %>%
  drop_na(tn:tp) # move an entire row when there's a NA
```

```
## 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.

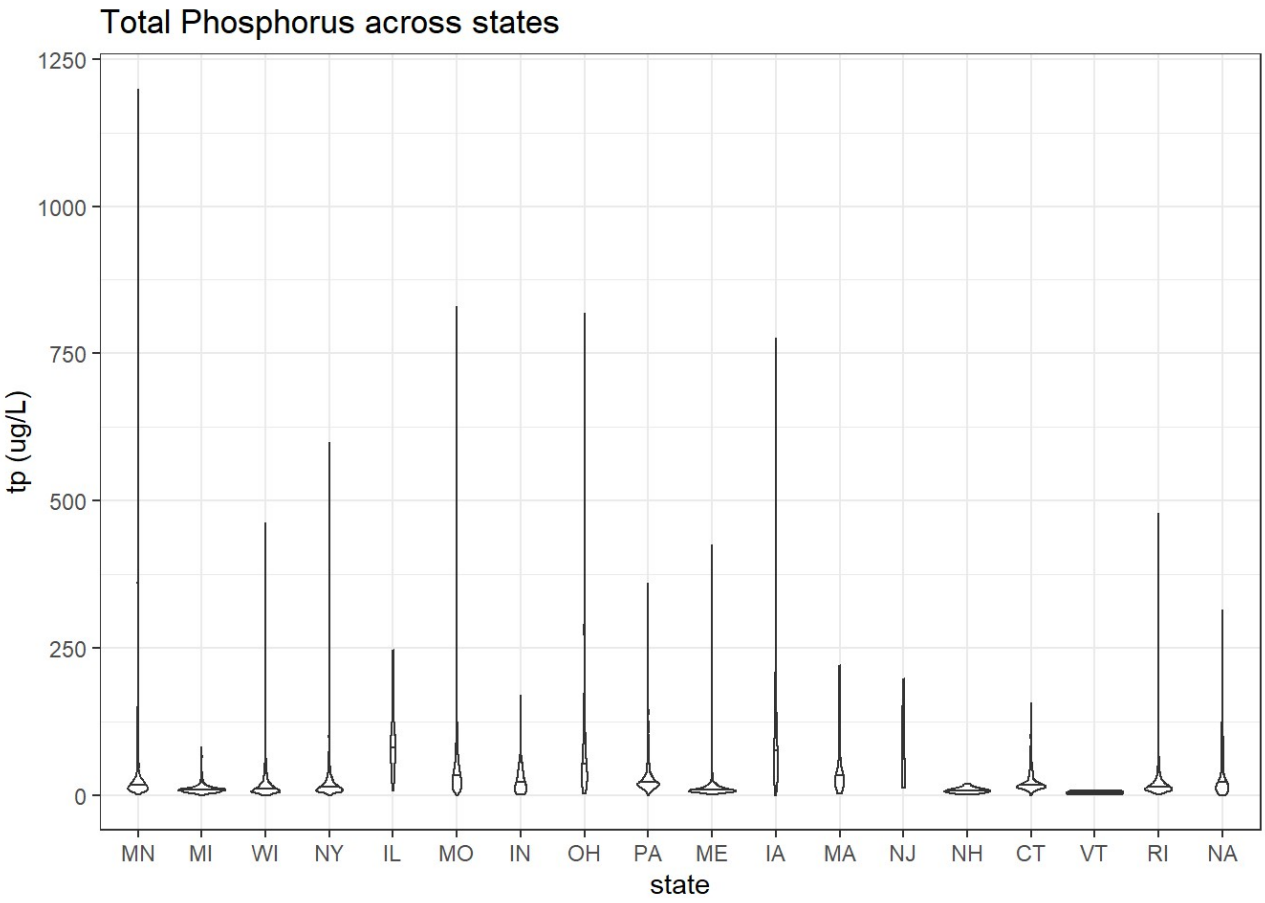
```
stateTNviolin <- ggplot(LAGOSNandP, aes(x = state, y = tn)) +
  geom_violin(draw_quantiles = 0.50) +
  labs(title = "Total Nitrogen across states", y = "tn (ug/L)")
print(stateTNviolin)
```

```
## 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

## 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) +
  labs(title = "Total Phosphorus across states", y = "tp (ug/L)")
print(stateTPviolin)
```

```
## 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

## 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
```




```
NitrogenMedian <- LAGOSNandP %>%
  group_by(state) %>%
  summarize(median = median(tn),range = max(tn)-min(tn))
```

```
## Warning: Factor `state` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

```
PhosphorusMedian <- LAGOSNandP %>%
  group_by(state) %>%
  summarize(median = median(tp), range = max(tp)-min(tp))
```

```
## Warning: Factor `state` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

Which states have the highest and lowest median concentrations?

> TN: Highest median: Iowa, Lowest median: Vermont

> TP: Highest median: Illinois, Lowest median: Vermont

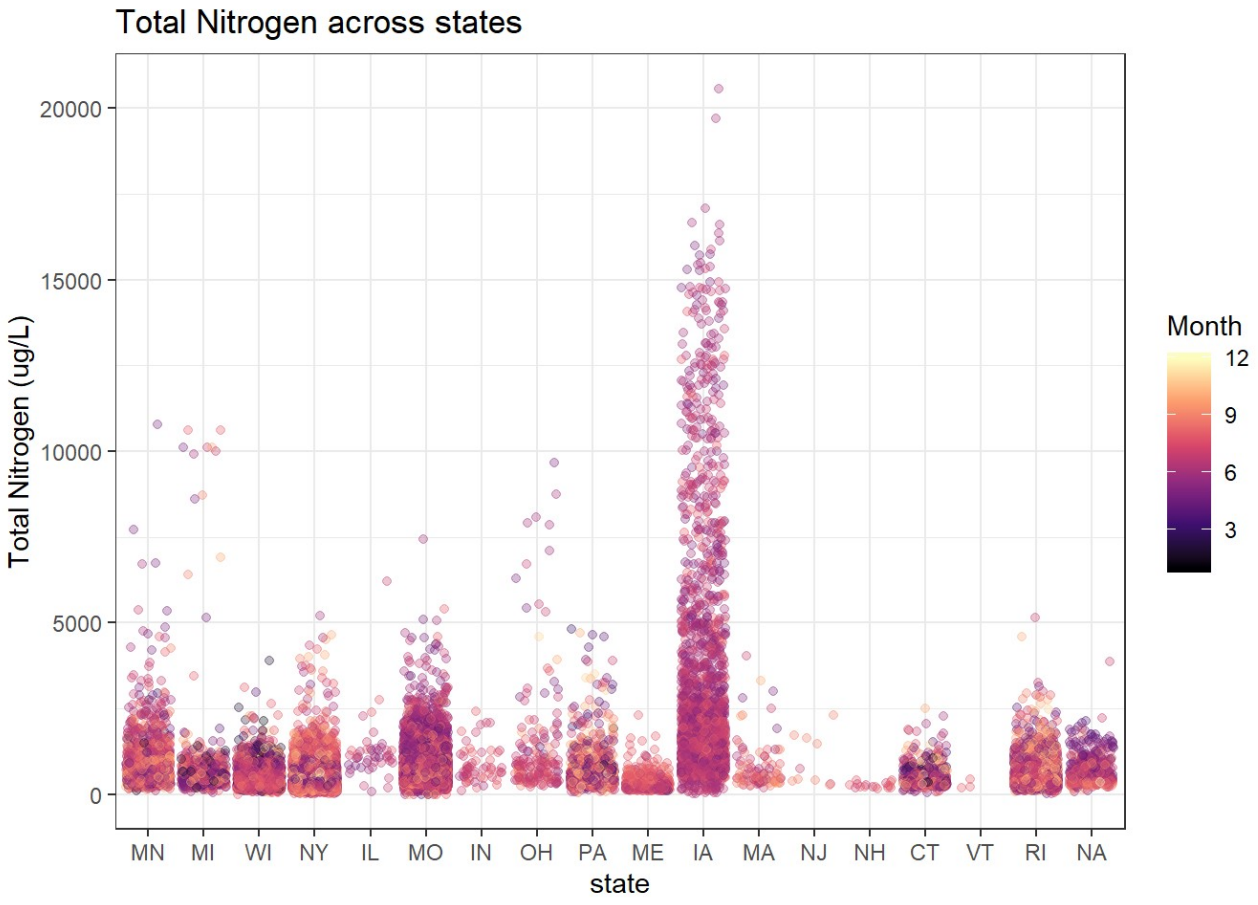
Which states have the highest and lowest concentration ranges?

> TN: Highest range: Iowa, Lowest range: Vermont

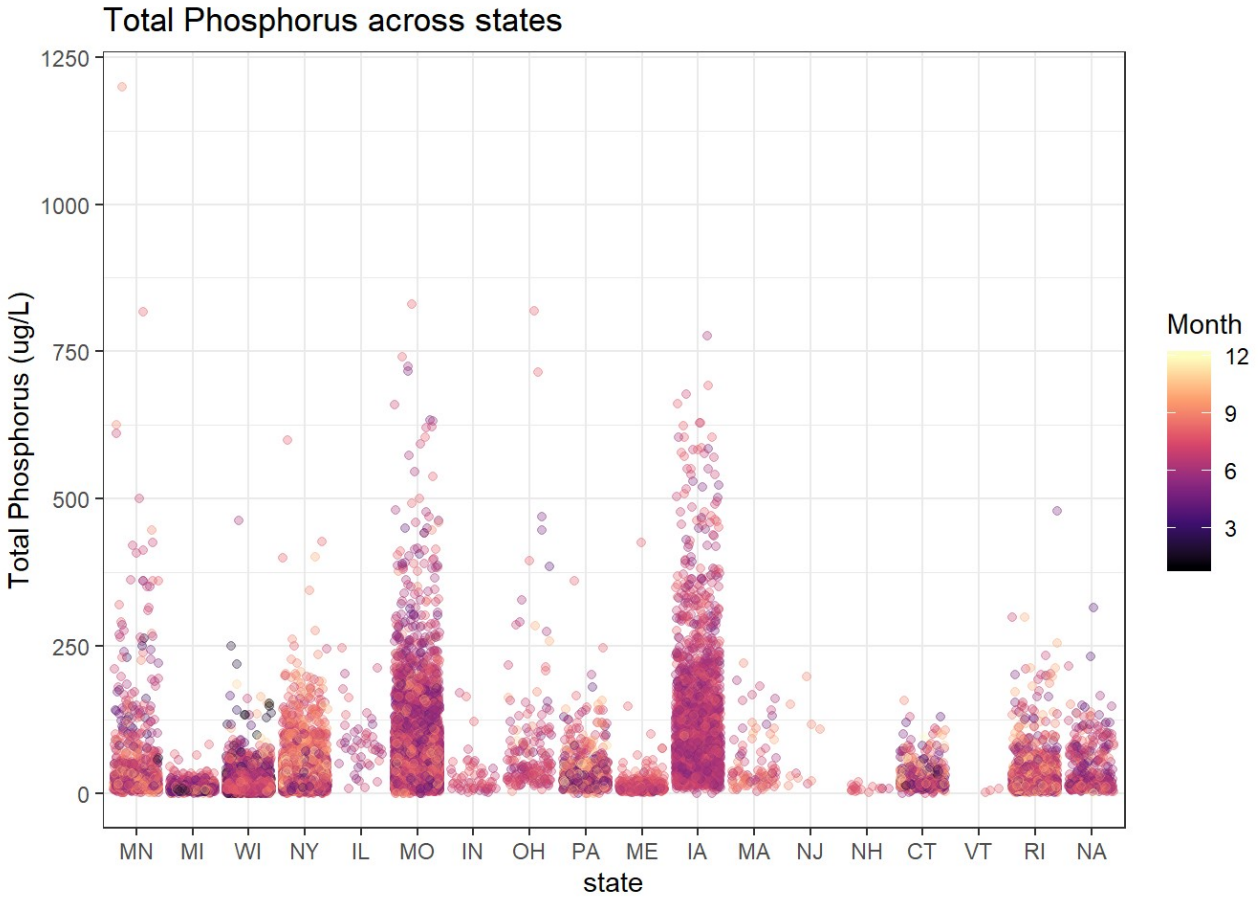
> TP: Highest range: Minnesota, Lowest range: Vermont

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(title = "Total Nitrogen across states", y = "Total Nitrogen (ug/L)", color = "M
onth") +
  theme(legend.position = "right") +
  scale_color_viridis_c(option = "magma")
print(stateTNjitter)
```



```
stateTPjitter <- ggplot(LAGOSNandP, aes(x = state, y = tp, color = samplemonth)) +
  geom_jitter(alpha = 0.3) +
  labs(title = "Total Phosphorus across states", y = "Total Phosphorus (ug/L)", color
= "Month") +
  theme(legend.position = "right") +
  scale_color_viridis_c(option = "magma")
print(stateTPjitter)
```




```
LAGOSNandP %>%
        group_by(state) %>%
        summarise(Number_of_Observations = n())
```

```
## Warning: Factor `state` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

state <fctr>	Number_of_Observations <int>
MN	1341
MI	877

state <fctr>	Number_of_Observations <int>
WI	2336
NY	7715
IL	46
MO	11412
IN	57
OH	166
PA	983
ME	633

1-10 of 18 rows Previous **1** 2 Next

```
LAGOSNandP %>%
        group_by(samplemonth, state) %>%
        summarise(Number_of_Observations = n())
```

```
## Warning: Factor `state` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

samplemonth <dbl>	state <fctr>	Number_of_Observations <int>
1	WI	146
1	MO	2
1	PA	10
1	CT	5
1	*NA*	1
2	MN	10
2	MI	11
2	WI	147
2	PA	13
2	CT	1

1-10 of 132 rows Previous **1** 2 3 4 5 6 … 14 Next

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

> TN: If we only look at the graph we may think that Iowa has the most samples, but if we count observations by state (since all rows with NA in EITHER tn OR tp have been dropped when wrangling with data, number of observations of tn should equal that of tp), we will find it's Missouri.

> TP: Missouri

> Impact: The state with the fewest observations (Vermont) has the lowest median and range, possibly just because not enough data are collected.

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

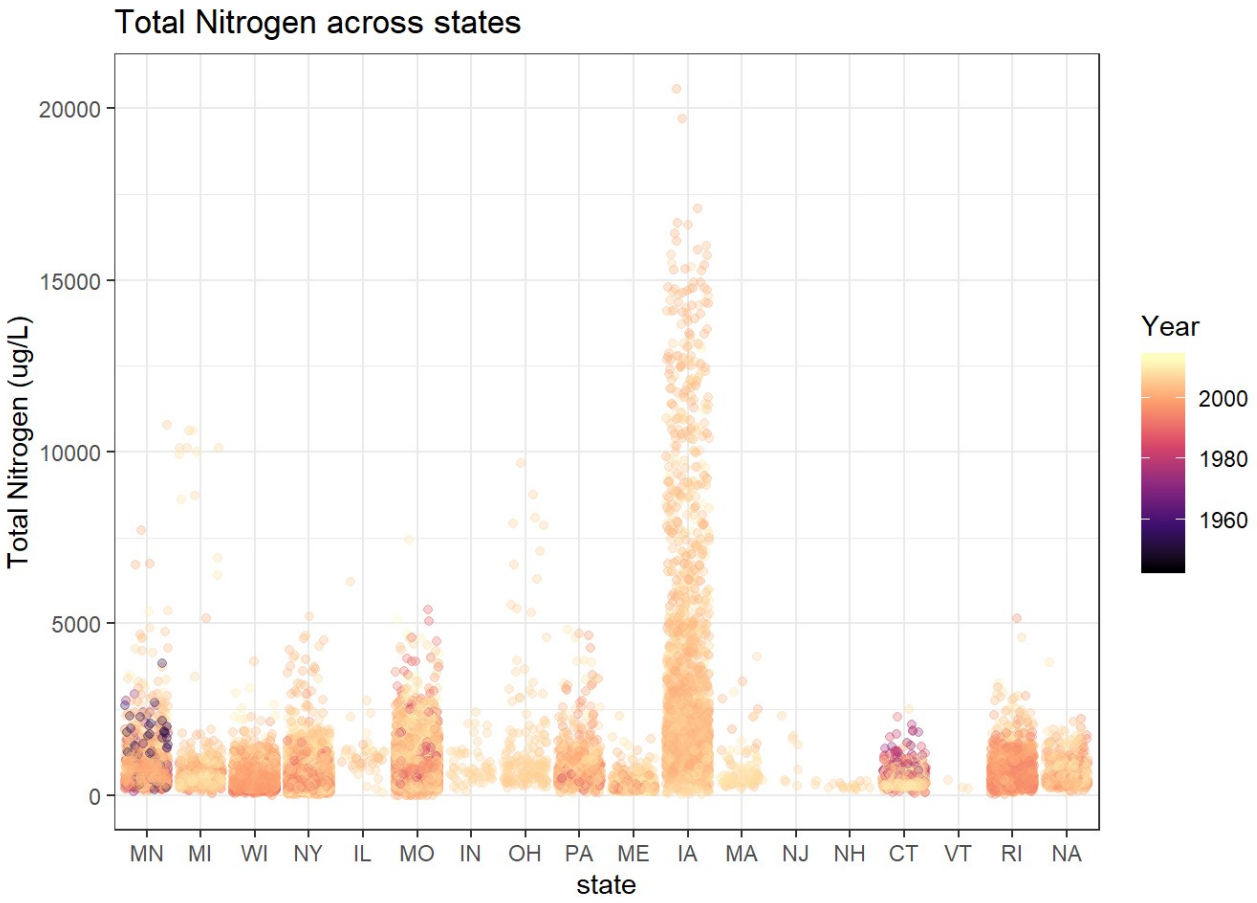
> TN: June, July, August

> TP: June, July, August

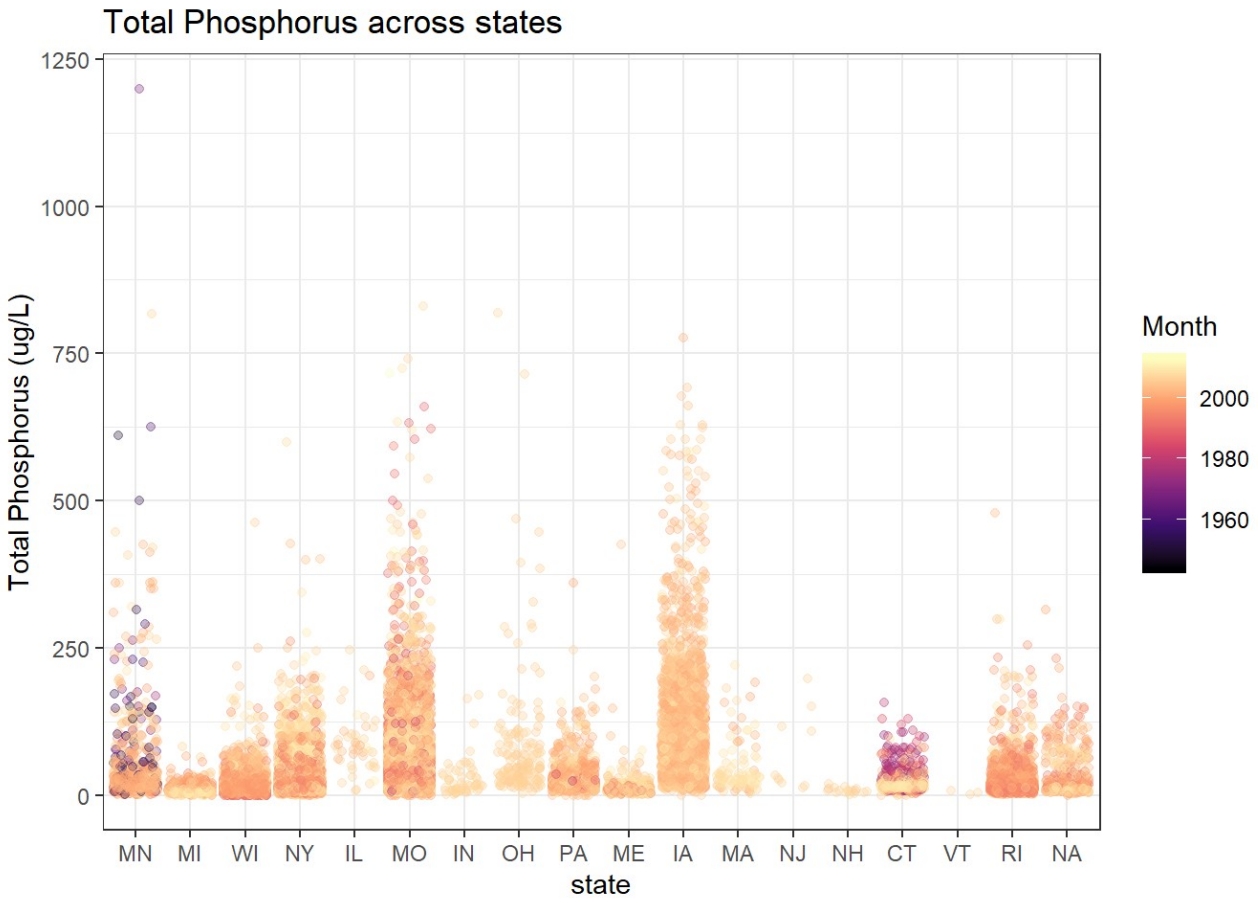
> It differs among states - for example, Wisconsin only samples in October, November and December.

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, y = tn, color = sampleyear)) +
  geom_jitter(alpha = 0.3) +
  labs(title = "Total Nitrogen across states", y = "Total Nitrogen (ug/L)", color = "Y
ear") +
  theme(legend.position = "right") +
  scale_color_viridis_c(option = "magma")
print(stateTNjitter.year)
```




```
stateTPjitter.year <- ggplot(LAGOSNandP, aes(x = state, y = tp, color = sampleyear)) +
  geom_jitter(alpha = 0.3) +
  labs(title = "Total Phosphorus across states", y = "Total Phosphorus (ug/L)", color
= "Month") +
  theme(legend.position = "right") +
  scale_color_viridis_c(option = "magma")
print(stateTPjitter.year)
```




```
LAGOSNandP %>%
        group_by(sampleyear, state) %>%
        summarise(Number_of_Observations = n())
```

```
## Warning: Factor `state` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

sampleyear	state	Number_of_Observations
<dbl>	<fctr>	<int>
1944	MN	1
1945	MN	2
1946	MN	3
1947	MN	2
1948	MN	2
1949	MN	5
1950	MN	1

sampleyear <dbl>	state <fctr>	Number_of_Observations <int>
1953	MN	8
1954	MN	6
1955	MN	14

1-10 of 306 rows Previous 1 2 3 4 5 6 … 31 Next

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

TN: 2007

TP: 2007

Yes. For example, for New York it is 2008 that is sampled most extensively, and for Missouri it is 2006.

Reflection

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

1. Using different metrics we would make different evaluations of a lake's eutriphication state.
2. Summer time is really the peak time of eutriphication.

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

The data on different tn, tp levels across months and eutriphication classes using different metrics.

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

Yes, it makes the process of exploration also part of the learning.

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

It is more messy - so many NAs!