Assignment 5: Water Quality in Lakes

Felipe Raby Amadori

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 pdf 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.

```
knitr::opts_chunk$set(message = FALSE, warning = FALSE)

#Verify your working directory is set to the R project file
getwd()
```

```
## [1] "C:/Users/Felipe/OneDrive - Duke University/1. DUKE/Ramos 3 Semestre/Hydrologic_Data_Analysis"
```

```
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(cowplot)
```

```
##
## ************************************************************
## Note: As of version 1.0.0, cowplot does not change the
##   default ggplot2 theme anymore. To recover the previous
```

```
##   behavior, execute:
##   theme_set(theme_cowplot())
```

```
## ******************************************************
```

```
library(LAGOSNE)
library(janitor)
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(lubridate)
```

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

```
## The following object is masked from 'package:cowplot':
##
##     stamp
```

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

```
#Set your ggplot theme (can be theme_classic or something else)
felipe_theme <- theme_light(base_size = 12) +
  theme(axis.text = element_text(color = "grey8"),
        legend.position = "right", plot.title = element_text(hjust = 0.5))
theme_set(felipe_theme)


#Load the LAGOSdata database and the trophic state index csv file we created on 2019/09/27.
load(file = "./Data/Raw/LAGOSdata.rda")
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).

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


LAGOStrophicComplete$trophic.class.secchi <-
  factor(LAGOStrophicComplete$trophic.class.secchi,
         levels = c("Oligotrophic", "Mesotrophic", "Eutrophic", "Hypereutrophic"))
```

```
LAGOStrophicComplete <-
  mutate(LAGOStrophicComplete,
         trophic.class.tp =
            ifelse(TSI.tp < 40, "Oligotrophic",
                   ifelse(TSI.tp < 50, "Mesotrophic",
                          ifelse(TSI.tp < 70, "Eutrophic", "Hypereutrophic"))))

LAGOStrophicComplete$trophic.class.tp <-
  factor(LAGOStrophicComplete$trophic.class.tp,
         levels = c("Oligotrophic", "Mesotrophic", "Eutrophic", "Hypereutrophic"))

head(LAGOStrophicComplete)
```

```
##   lagoslakeid sampledate   chla  tp secchi    gnis_name lake_area_ha state
## 1      126841 1985-11-05   8.40 100 1.0000  Silver Lake     32.38997    NY
## 2        6456 2006-06-15   5.64  14 2.2098         <NA>    861.58209    IL
## 3        6469 2006-06-19  41.90  88 0.4572 Tampier Lake     48.97546    IL
## 4       81320 1985-11-07   1.40 100 0.5000         <NA>     11.05260    NY
## 5      122514 1988-06-23 133.20 700 0.4000   Evens Lake     11.24954    NY
## 6        6450 2006-08-23  12.30  14 0.9144         <NA>    110.85536    IL
##   state_name sampleyear samplemonth season TSI.chl TSI.secchi TSI.tp
## 1   New York       1985          11   Fall      60         60     71
## 2   Illinois       2006           6 Summer      57         49     42
## 3   Illinois       2006           6 Summer      76         71     69
## 4   New York       1985          11   Fall      43         70     71
## 5   New York       1988           6 Summer      88         73     99
## 6   Illinois       2006           8 Summer      64         61     42
##    trophic.class trophic.class.secchi trophic.class.tp
## 1      Eutrophic            Eutrophic   Hypereutrophic
## 2      Eutrophic          Mesotrophic      Mesotrophic
## 3 Hypereutrophic       Hypereutrophic        Eutrophic
## 4    Mesotrophic       Hypereutrophic   Hypereutrophic
## 5 Hypereutrophic       Hypereutrophic   Hypereutrophic
## 6      Eutrophic            Eutrophic      Mesotrophic
```

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.

```
#Number of observations that fall into the four trophic state categories for trophic.class
#(using chlorophyll a concentration)

N_Obs_TrophicClass_Chl <- data.frame(summary(LAGOStrophicComplete$trophic.class))
names(N_Obs_TrophicClass_Chl)[1] <- c("Number_Observations")
N_Obs_TrophicClass_Chl <- data.frame(TrophicState = rownames(N_Obs_TrophicClass_Chl),
                                     N_Obs_TrophicClass_Chl)
rownames(N_Obs_TrophicClass_Chl) <- c()
N_Obs_TrophicClass_Chl <- adorn_totals(N_Obs_TrophicClass_Chl,"row")

#The number of observations are presented in the following dataframe
N_Obs_TrophicClass_Chl
```

```
##    TrophicState Number_Observations
##       Eutrophic               41861
##  Hypereutrophic               14379
##     Mesotrophic               15413
```

```
##    Oligotrophic               3298
##           Total              74951
```

```
#Number of observations that fall into the four trophic state categories for
#trophic.class.secchi (using Secchi Depth)

N_Obs_TrophicClass_Secchi <- data.frame(summary(LAGOStrophicComplete$trophic.class.secchi))
names(N_Obs_TrophicClass_Secchi)[1] <- c("Number_Observations")
N_Obs_TrophicClass_Secchi <-
  data.frame(TrophicState = rownames(N_Obs_TrophicClass_Secchi),
             N_Obs_TrophicClass_Secchi)
rownames(N_Obs_TrophicClass_Secchi) <- c()
N_Obs_TrophicClass_Secchi <- adorn_totals(N_Obs_TrophicClass_Secchi,"row")

#The number of observations are presented in the following dataframe
N_Obs_TrophicClass_Secchi
```

```
##    TrophicState Number_Observations
##    Oligotrophic               16110
##     Mesotrophic               25083
##       Eutrophic               28659
##  Hypereutrophic                5099
##           Total              74951
```

```
#Number of observations that fall into the four trophic state categories for
#trophic.class.tp (using Total phosphorus)

N_Obs_TrophicClass_tp <- data.frame(summary(LAGOStrophicComplete$trophic.class.tp))
names(N_Obs_TrophicClass_tp)[1] <- c("Number_Observations")
N_Obs_TrophicClass_tp <- data.frame(TrophicState = rownames(N_Obs_TrophicClass_tp),
                                    N_Obs_TrophicClass_tp)
rownames(N_Obs_TrophicClass_tp) <- c()
N_Obs_TrophicClass_tp <- adorn_totals(N_Obs_TrophicClass_tp,"row")

#The number of observations are presented in the following dataframe
N_Obs_TrophicClass_tp
```

```
##    TrophicState Number_Observations
##    Oligotrophic               19861
##     Mesotrophic               23023
##       Eutrophic               24839
##  Hypereutrophic                7228
##           Total              74951
```

7. What proportion of total observations are considered eutrophic or hypereutrophic according to the three different metrics (trophic.class, trophic.class.secchi, trophic.class.tp)?

```
#proportion of total observations of the four trophic state categories for
#chlorophyll a concentration metric

N_Obs_TrophicClass_Chl <-
  mutate(N_Obs_TrophicClass_Chl,
         Proportion = N_Obs_TrophicClass_Chl$Number_Observations/N_Obs_TrophicClass_Chl[5,2])

#The proportion of observations is presented in the folowing dataframe
N_Obs_TrophicClass_Chl
```

```
##     TrophicState Number_Observations Proportion
## 1      Eutrophic               41861 0.55851156
## 2 Hypereutrophic               14379 0.19184534
## 3    Mesotrophic               15413 0.20564102
## 4   Oligotrophic                3298 0.04400208
## 5          Total               74951 1.00000000
```

20.6% + 4.4% = 25% of total observations are considered eutrophic or hypereutrophic for chlorophyll a concentration metric

```
#proportion of total observations of the four trophic state categories for Secchi Depth metric

N_Obs_TrophicClass_Secchi <-
  mutate(N_Obs_TrophicClass_Secchi,
         Proportion = N_Obs_TrophicClass_Secchi$Number_Observations/N_Obs_TrophicClass_Secchi[5,2])

#The proportion of observations is presented in the folowing dataframe
N_Obs_TrophicClass_Secchi
```

```
##     TrophicState Number_Observations Proportion
## 1   Oligotrophic               16110 0.21494043
## 2    Mesotrophic               25083 0.33465864
## 3      Eutrophic               28659 0.38236981
## 4 Hypereutrophic                5099 0.06803111
## 5          Total               74951 1.00000000
```

38.2% + 6.8% = 45% of total observations are considered eutrophic or hypereutrophic for Secchi Depth metric

```
#proportion of total observations of the four trophic state categories for Total phosphorus metric

N_Obs_TrophicClass_tp <-
  mutate(N_Obs_TrophicClass_tp,
         Proportion = N_Obs_TrophicClass_tp$Number_Observations/N_Obs_TrophicClass_tp[5,2])

#The proportion of observations is presented in the folowing dataframe
N_Obs_TrophicClass_tp
```

```
##     TrophicState Number_Observations Proportion
## 1   Oligotrophic               19861 0.26498646
## 2    Mesotrophic               23023 0.30717402
## 3      Eutrophic               24839 0.33140318
## 4 Hypereutrophic                7228 0.09643634
## 5          Total               74951 1.00000000
```

33.1% + 9.6% = 42.7% of total observations are considered eutrophic or hypereutrophic for Total phosphorus metric

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

With 45% of total observations considered euthrophic or hypereutrophic compared to 25% and 42.7%, the Secchi Depth metric is the most conservative. It is probably because is the least accurate estimation. I measures water transparency to indirectly give an estimation of concentration of suspended and dissolved material in the water, which then is used to derive the biomass. In some cases water transparency could be altered not only by biomass, which would lead to overestimation of biomass by the Secchi disk metric and therefore overestimation of the trophic state.

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.

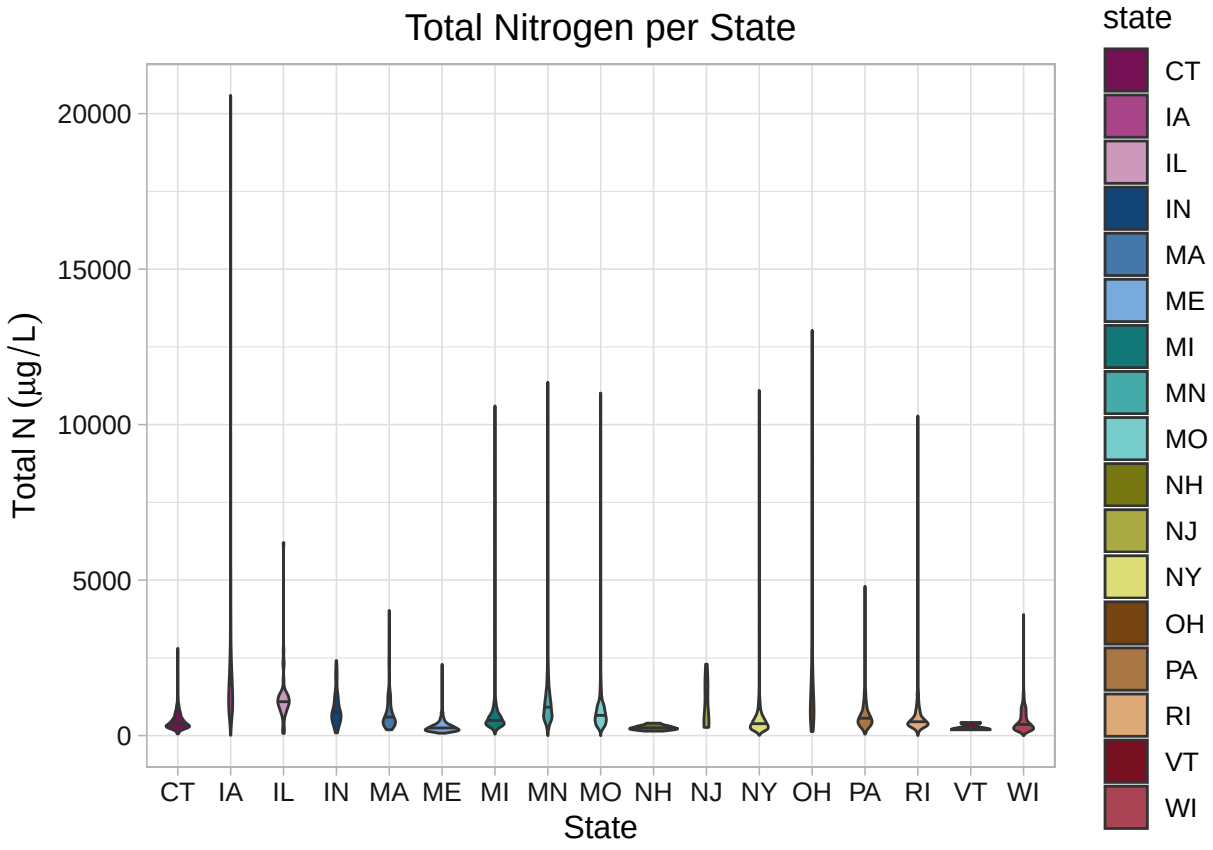
```
LAGOSlocus <- LAGOSdata$locus
LAGOSstate <- LAGOSdata$state
LAGOSnutrient <- LAGOSdata$epi_nutr


# 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 data frames
LAGOSlocations <- left_join(LAGOSlocus, LAGOSstate, by = "state_zoneid")
LAGOSNandP <- left_join(LAGOSnutrient, LAGOSlocations, by = "lagoslakeid")


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

head(LAGOSNandP)
```

```
##   lagoslakeid sampledate tn  tp state state_name sampleyear samplemonth
## 1      126841 1985-11-05 NA 100    NY   New York       1985          11
## 2        6456 2006-06-15 NA  14    IL   Illinois       2006           6
## 3        6469 2006-06-19 NA  88    IL   Illinois       2006           6
## 4       81320 1985-11-07 NA 100    NY   New York       1985          11
## 5      122514 1988-06-23 NA 700    NY   New York       1988           6
## 6        6450 2006-08-23 NA  14    IL   Illinois       2006           8
```

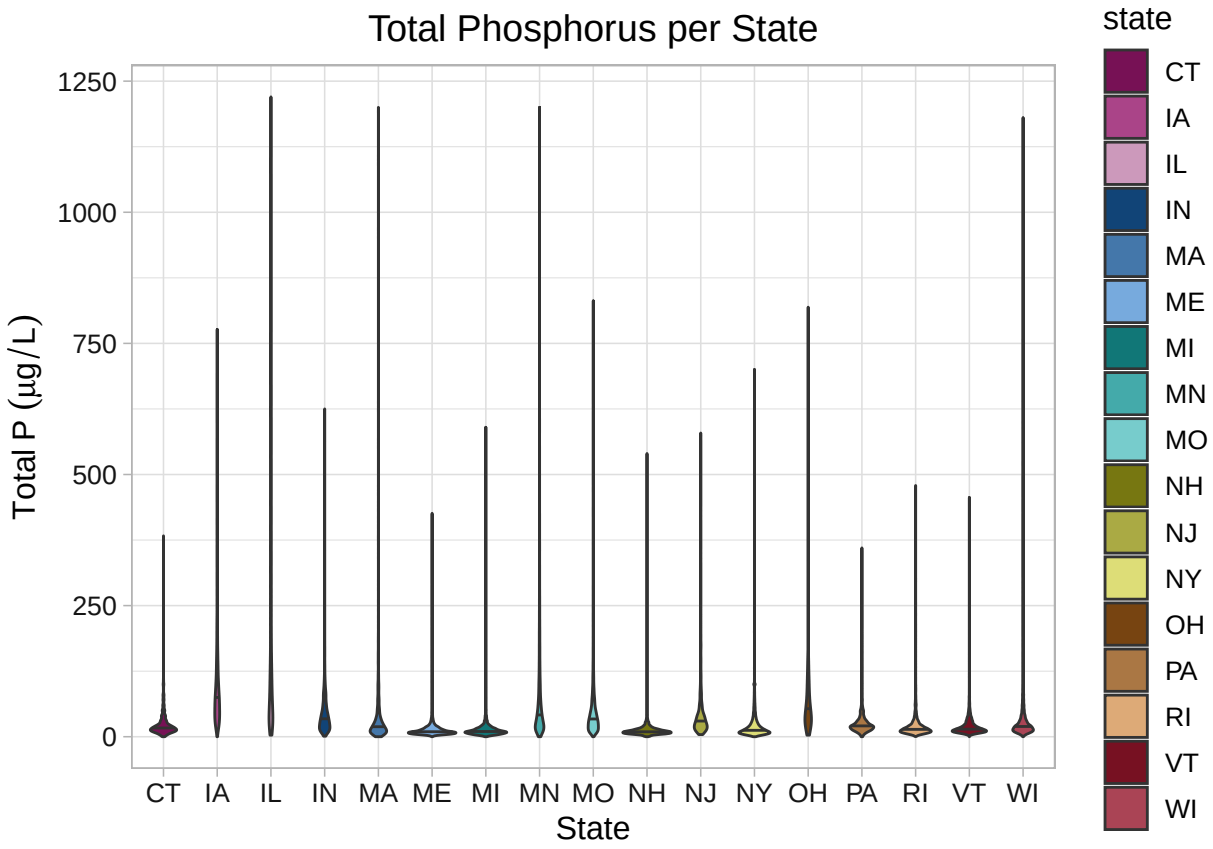
9. Create two violin plots comparing TN and TP concentrations across states. Include a 50th percentile line inside the violins.

```
LAGOSN <-
  LAGOSNandP %>%
  drop_na(tn,state)

LAGOSP <-
  LAGOSNandP %>%
  drop_na(tp,state)


LAGOSNviolin <- ggplot(LAGOSN, aes(x =state, fill =state)) +
  geom_violin(aes(y = tn), draw_quantiles = 0.50) +
  scale_fill_manual(values = c("#771155", "#AA4488", "#CC99BB", "#114477", "#4477AA",
                               "#77AADD", "#117777", "#44AAAA", "#77CCCC", "#777711",
                               "#AAAA44", "#DDDD77", "#774411", "#AA7744", "#DDAA77",
```

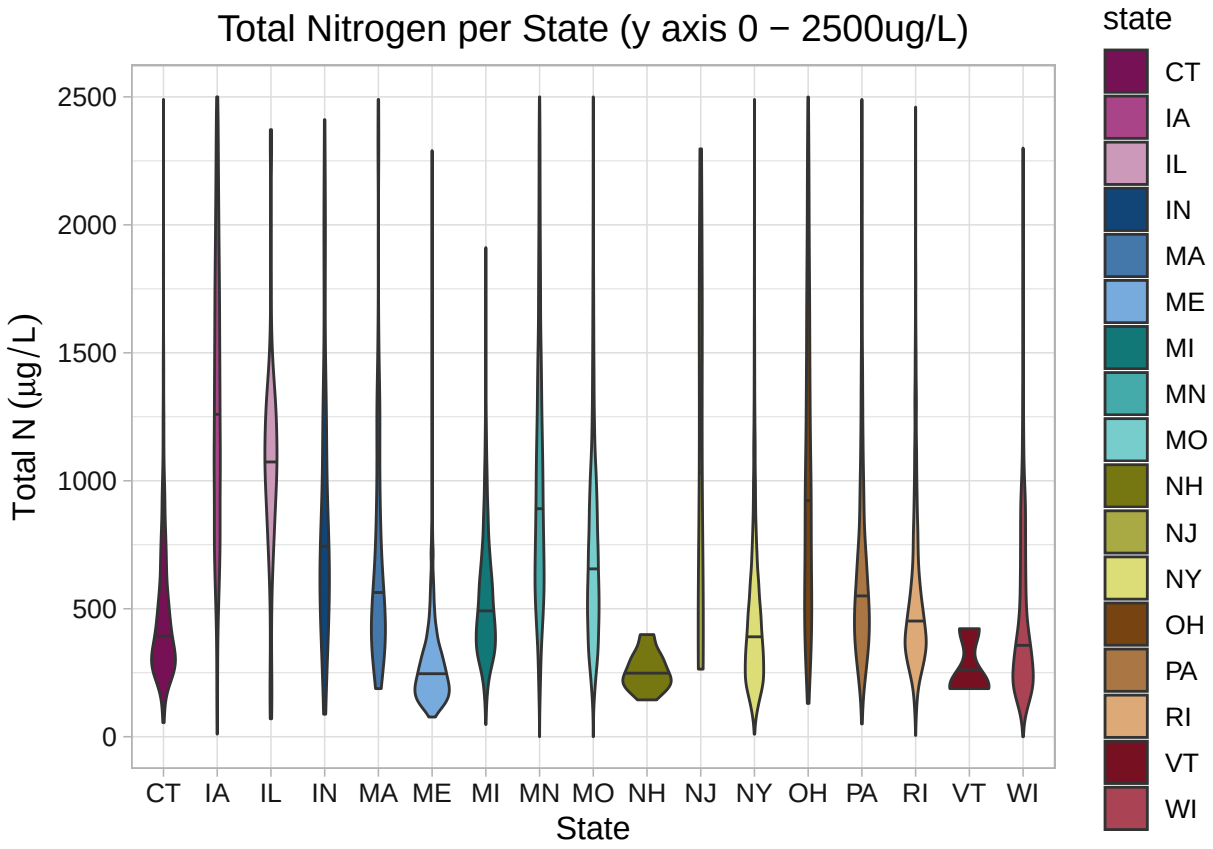
```
                                "#771122", "#AA4455")) +
  xlab("State") +
  ylab(Total ~ N ~ (mu*g / L)) +
  ggtitle("Total Nitrogen per State")
print(LAGOSNviolin)
```



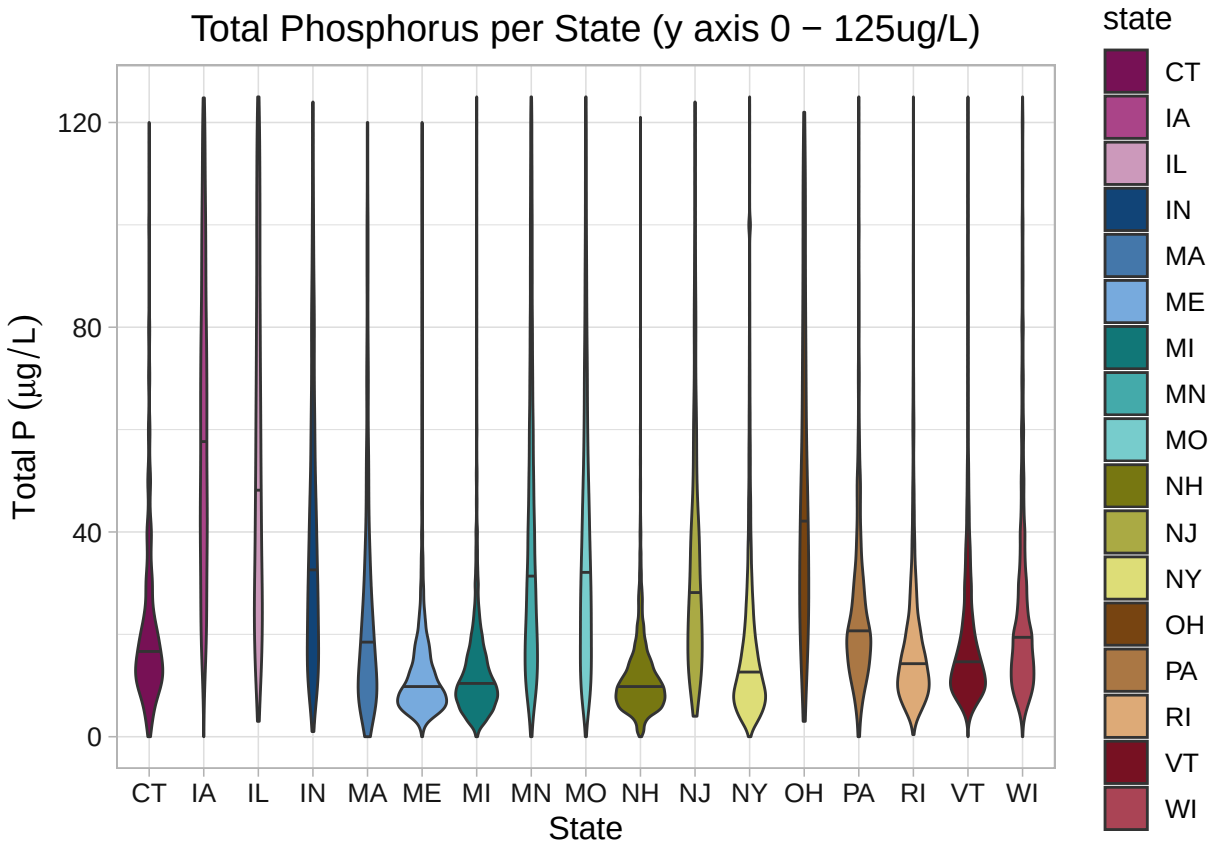

```
LAGOSPviolin <- ggplot(LAGOSP, aes(x =state, fill = state)) +
  geom_violin(aes(y = tp), draw_quantiles = 0.50) +
    scale_fill_manual(values = c("#771155", "#AA4488", "#CC99BB", "#114477", "#4477AA",
                                 "#77AADD", "#117777", "#44AAAA", "#77CCCC", "#777711",
                                 "#AAAA44", "#DDDD77", "#774411", "#AA7744", "#DDAA77",
                                 "#771122", "#AA4455")) +
  xlab("State") +
  ylab(Total ~ P ~ (mu*g / L)) +
  ggtitle("Total Phosphorus per State")
print(LAGOSPviolin)
```




```
# To better see the median, I scaled the y axis

LAGOSNviolin2 <- ggplot(LAGOSN, aes(x =state, fill =state)) +
  geom_violin(aes(y = tn), draw_quantiles = 0.50) +
  ylim(0,2500) +
  scale_fill_manual(values = c("#771155", "#AA4488", "#CC99BB", "#114477", "#4477AA",
                               "#77AADD", "#117777", "#44AAAA", "#77CCCC", "#777711",
                               "#AAAA44", "#DDDD77", "#774411", "#AA7744", "#DDAA77",
                               "#771122", "#AA4455")) +
  xlab("State") +
  ylab(Total ~ N ~ (mu*g / L)) +
  ggtitle("Total Nitrogen per State (y axis 0 - 2500ug/L)")
print(LAGOSNviolin2)
```




```
LAGOSPviolin2 <- ggplot(LAGOSP, aes(x =state, fill = state)) +
  geom_violin(aes(y = tp), draw_quantiles = 0.50) +
  ylim(0,125) +
    scale_fill_manual(values = c("#771155", "#AA4488", "#CC99BB", "#114477", "#4477AA",
                                 "#77AADD", "#117777", "#44AAAA", "#77CCCC", "#777711",
                                 "#AAAA44", "#DDDD77", "#774411", "#AA7744", "#DDAA77",
                                 "#771122", "#AA4455")) +
  xlab("State") +
  ylab(Total ~ P ~ (mu*g / L)) +
  ggtitle("Total Phosphorus per State (y axis 0 - 125ug/L)")
print(LAGOSPviolin2)
```




```
#To be sure about the next questions

LAGOSN_Summary <- LAGOSN %>%
  group_by(state_name) %>%
  summarize(tN_Median = median(tn),
            tN_Range = max(tn) - min(tn))

LAGOSN_Summary
```

```
## # A tibble: 17 x 3
##    state_name    tN_Median tN_Range
##    <chr>             <dbl>    <dbl>
##  1 Connecticut        390     2751
##  2 Illinois          1084.    6133
##  3 Indiana            714     2323
##  4 Iowa              1628.   20564.
##  5 Maine              246.    2213
##  6 Massachusetts      532     3834
##  7 Michigan           493    10552.
##  8 Minnesota          920    11350
##  9 Missouri           660    11010
## 10 New Hampshire      244      255
## 11 New Jersey         585     2033
## 12 New York           390    11090
## 13 Ohio              1343    12889.
## 14 Pennsylvania       550     4750
```

```
## 15 Rhode Island       450    10275
## 16 Vermont            204      234
## 17 Wisconsin          352     3893
```

```
LAGOSP_Summary <- LAGOSP %>%
  group_by(state_name) %>%
  summarize(tP_Median = median(tp),
            tP_Range = max(tp) - min(tp))

LAGOSP_Summary
```

```
## # A tibble: 17 x 3
##    state_name    tP_Median tP_Range
##    <chr>             <dbl>    <dbl>
##  1 Connecticut        16.4     383
##  2 Illinois           83.1    1217
##  3 Indiana            34       624
##  4 Iowa               74.4     776.
##  5 Maine              10       426
##  6 Massachusetts      19      1200
##  7 Michigan           10.3     590
##  8 Minnesota          40      1200
##  9 Missouri           34       831
## 10 New Hampshire      10       540
## 11 New Jersey         29       575.
## 12 New York           13       700
## 13 Ohio               47.1     816
## 14 Pennsylvania       20       360
## 15 Rhode Island       14       479.
## 16 Vermont            14.9     456
## 17 Wisconsin          20      1180
```

Which states have the highest and lowest median concentrations?

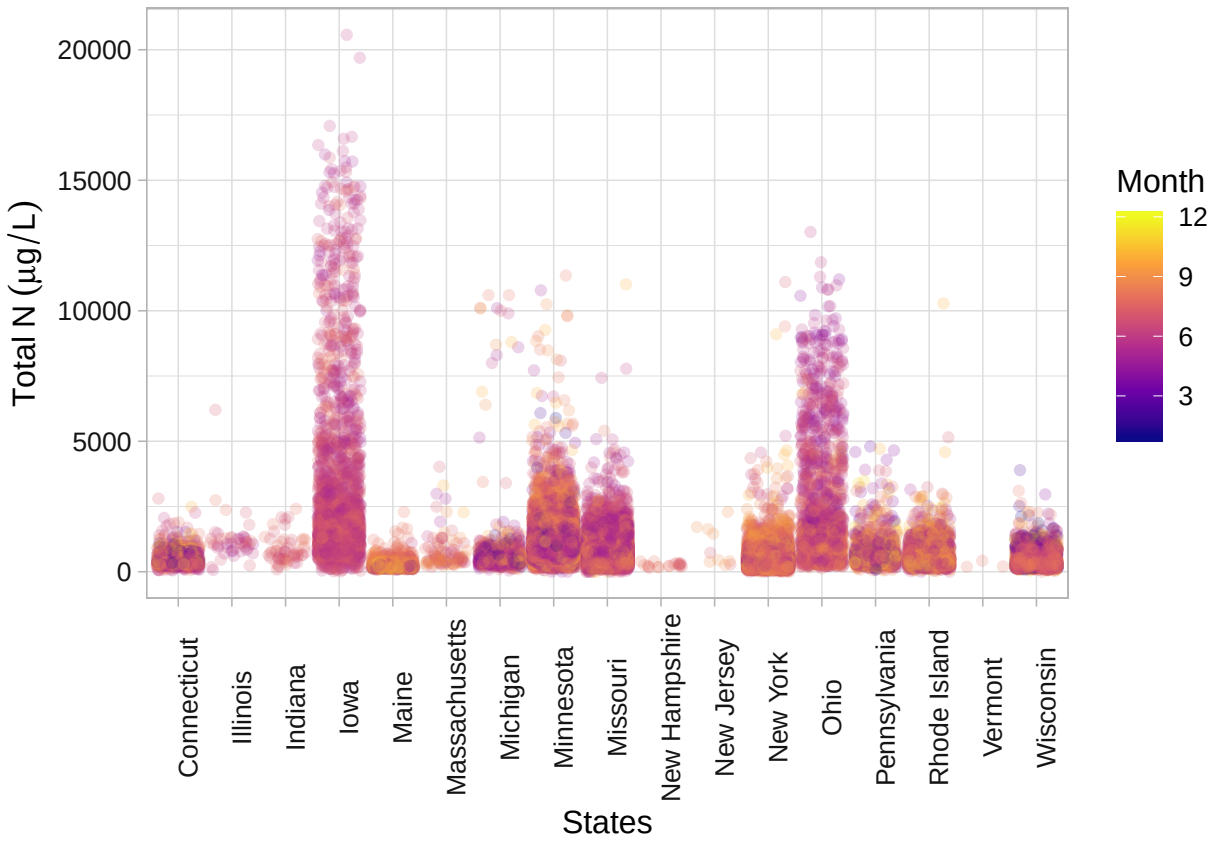
> TN: Highest: Iowa. Lowest: Vermont. I couldn't figure out why the Total N plot shows a higher median for Vermont (over 250 ug/L)
>
> TP: Highest: Illinois. Lowest: New Hampshire.

Which states have the highest and lowest concentration ranges?

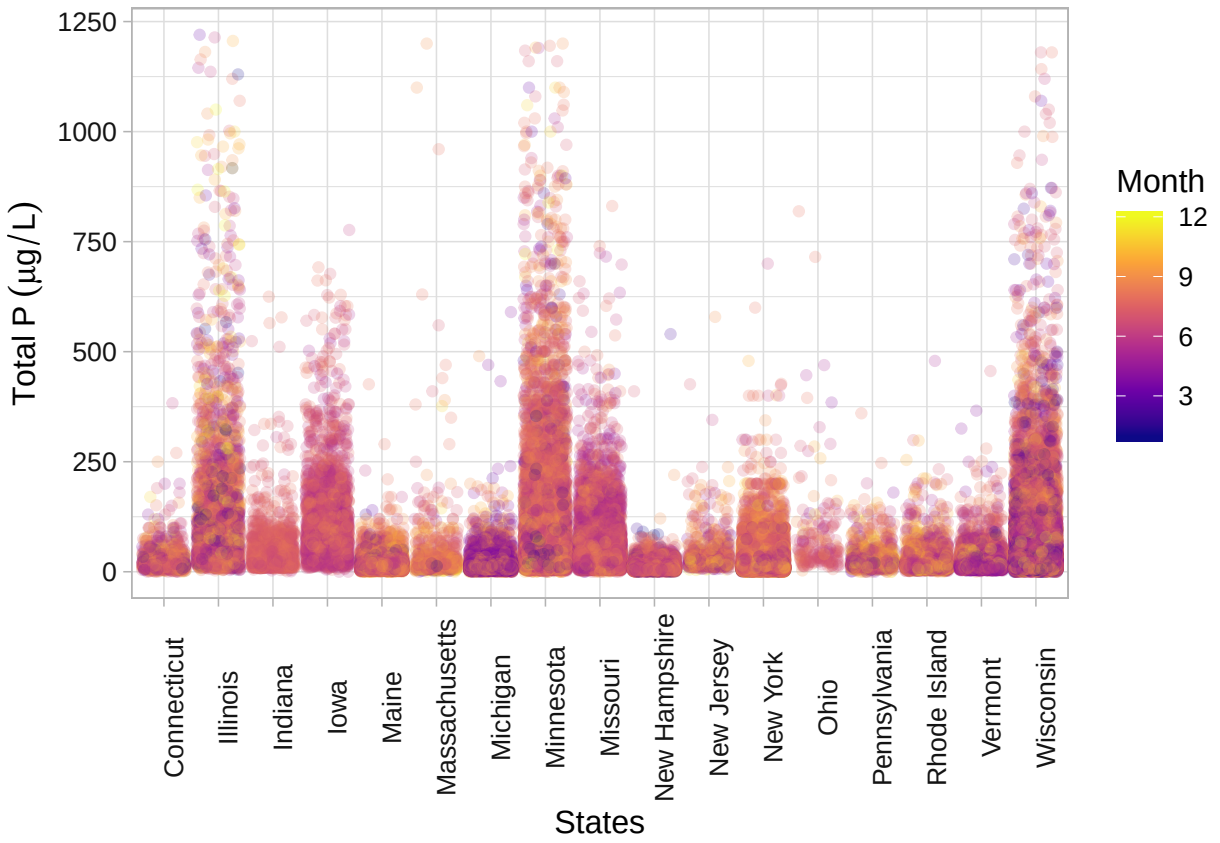
> TN: Highest: Iowa. Lowest: Vermont
>
> TP: Highest: Illinois. Lowest: Connecticut

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.

```
tnbystate <-
ggplot(LAGOSN,
       aes(x = as.factor(state_name), y = tn, color = samplemonth)) +
  geom_jitter(alpha = 0.2) +
  labs(x = "States", y = expression(Total ~ N ~ (mu*g / L)), color = "Month") +
  scale_color_viridis_c(option = "plasma") +
  theme(axis.text.x = element_text(angle = 90))
print(tnbystate)
```




```
tpbystate <-
ggplot(LAGOSP,
       aes(x = as.factor(state_name), y = tp, color = samplemonth)) +
  geom_jitter(alpha = 0.2) +
  labs(x = "States", y = expression(Total ~ P ~ (mu*g / L)), color = "Month") +
  scale_color_viridis_c(option = "plasma") +
  theme(axis.text.x = element_text(angle = 90))
print(tpbystate)
```




```
LAGOSN_Summary2 <- LAGOSN %>%
  group_by(state_name) %>%
  summarize(tN_Samples = length(tn))

LAGOSN_Summary2
```

```
## # A tibble: 17 x 2
##    state_name    tN_Samples
##    <chr>              <int>
##  1 Connecticut          916
##  2 Illinois              46
##  3 Indiana               57
##  4 Iowa                2649
##  5 Maine                762
##  6 Massachusetts         95
##  7 Michigan             885
##  8 Minnesota           8604
##  9 Missouri           11503
## 10 New Hampshire         19
## 11 New Jersey            10
## 12 New York            8091
## 13 Ohio                1502
## 14 Pennsylvania        1044
## 15 Rhode Island        2836
## 16 Vermont                3
## 17 Wisconsin           2416
```

```
LAGOSP_Summary2 <- LAGOSP %>%
  group_by(state_name) %>%
  summarize(tP_Samples = length(tp))

LAGOSP_Summary2
```

```
## # A tibble: 17 x 2
##    state_name    tP_Samples
##    <chr>              <int>
##  1 Connecticut         1222
##  2 Illinois            2632
##  3 Indiana             1340
##  4 Iowa                2920
##  5 Maine              11987
##  6 Massachusetts        657
##  7 Michigan           10250
##  8 Minnesota          11186
##  9 Missouri           11786
## 10 New Hampshire       8164
## 11 New Jersey           516
## 12 New York           21343
## 13 Ohio                 175
## 14 Pennsylvania        1240
## 15 Rhode Island        3612
## 16 Vermont             7980
## 17 Wisconsin          45743
```

```
LAGOSN_Summary3 <- LAGOSN %>%
  group_by(samplemonth) %>%
  summarize(tN_Samples = length(tn))

LAGOSN_Summary3
```

```
## # A tibble: 12 x 2
##    samplemonth tN_Samples
##          <dbl>      <int>
##  1           1        191
##  2           2        267
##  3           3        194
##  4           4       1507
##  5           5       5173
##  6           6       8014
##  7           7       9880
##  8           8       8451
##  9           9       4344
## 10          10       2803
## 11          11        537
## 12          12         77
```

```
LAGOSP_Summary3 <- LAGOSP %>%
  group_by(samplemonth) %>%
  summarize(tP_Samples = length(tp))

LAGOSP_Summary3
```

```
## # A tibble: 12 x 2
##    samplemonth tP_Samples
##          <dbl>      <int>
##  1           1       1279
##  2           2       2215
##  3           3       1821
##  4           4       8790
##  5           5      14221
##  6           6      22637
##  7           7      30410
##  8           8      31606
##  9           9      16094
## 10          10      10250
## 11          11       2826
## 12          12        604
```

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

TN: Missouri 11.503, Minnesota 8604, and New York 8.091

TP: Wisconsin 45.743 and New York 21.343

Number of samples totally could impact the total ranges of samples because a higher number of samples probably means that the samples were taken at different seasons, times, environmental conditions, and for a longer period of time (which makes more probable the presence of outliers) than data sets with fewer samples.

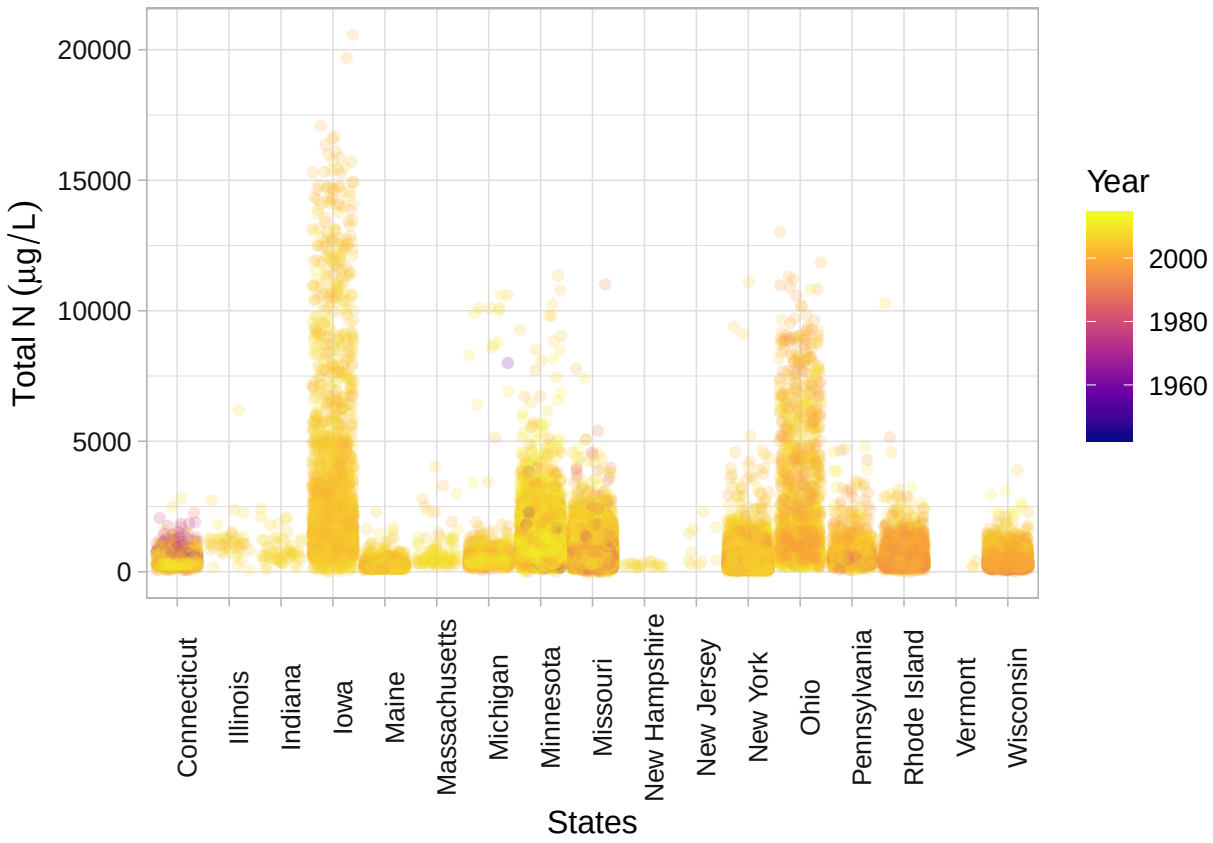
Which months are sampled most extensively? Does this differ among states?

TN: July, August, and June in decreasing order. According to the jitter plot, it differ among states. For example it can be seen that Iowa has more red/pink/purple values (summer months) and New York, Maine, and Rhode Island have a greater concentrations of yellow/orange values (fall months).

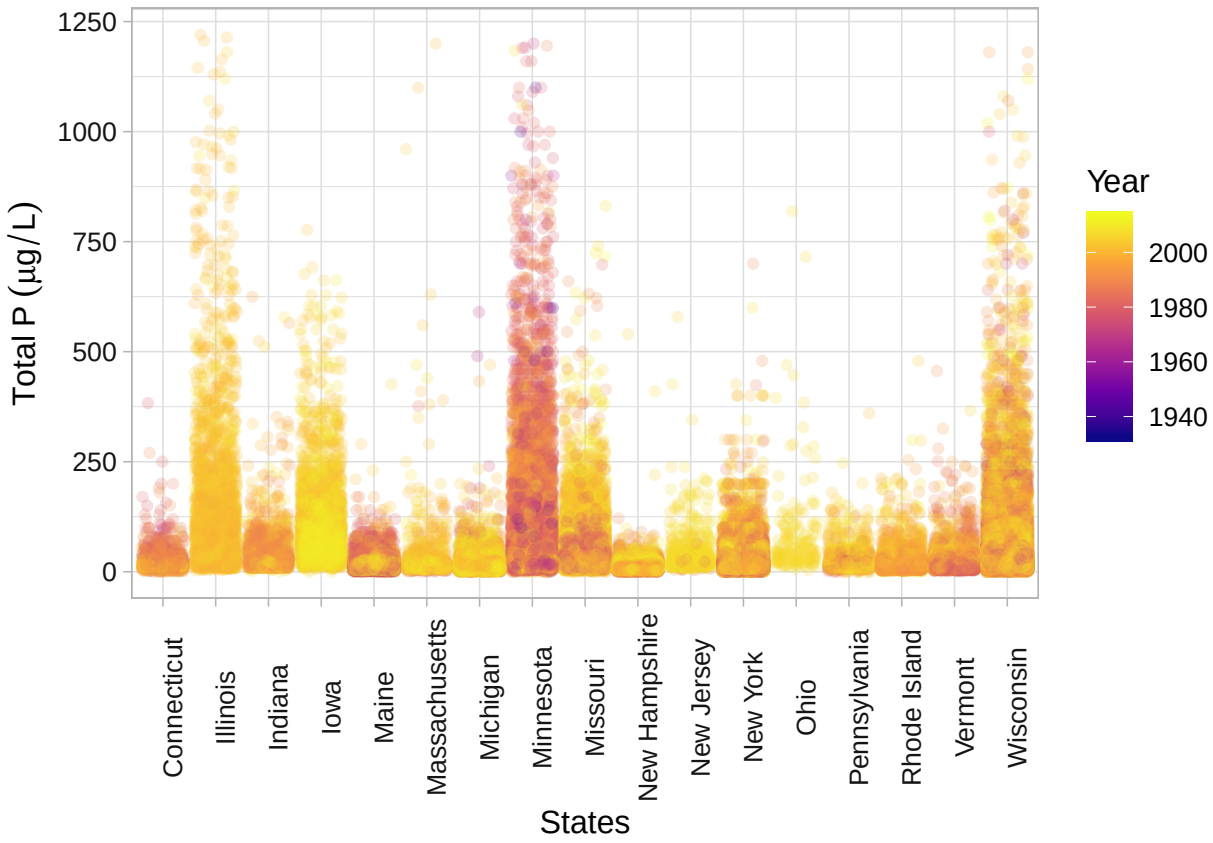
TP: August, July, and June in decreasing order. According to the jitter plot, it differ among states. For example it can be seen that Michigan, Vermont and Wisconsin have more blue/purple values (winter months) than the rest of the states. New York, Massachusetts, and Rhode Island have a greater concentrations of yellow/orange values (fall 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.

```
tnbystate <-
ggplot(LAGOSN,
       aes(x = as.factor(state_name), y = tn, color = sampleyear)) +
  geom_jitter(alpha = 0.2) +
  labs(x = "States", y = expression(Total ~ N ~ (mu*g / L)), color = "Year") +
  scale_color_viridis_c(option = "plasma") +
  theme(axis.text.x = element_text(angle = 90))
print(tnbystate)
```




```
tpbystate <-
ggplot(LAGOSP,
       aes(x = as.factor(state_name), y = tp, color = sampleyear)) +
  geom_jitter(alpha = 0.2) +
  labs(x = "States", y = expression(Total ~ P ~ (mu*g / L)), color = "Year") +
  scale_color_viridis_c(option = "plasma") +
  theme(axis.text.x = element_text(angle = 90))
print(tpbystate)
```




```
LAGOSN_Summary4 <- LAGOSN %>%
  group_by(sampleyear) %>%
  summarize(tN_Samples = length(tn))

tail(LAGOSN_Summary4, n=15)
```

```
## # A tibble: 15 x 2
##    sampleyear tN_Samples
##         <dbl>      <int>
##  1       1999       1242
##  2       2000       1119
##  3       2001       1863
##  4       2002       2253
##  5       2003       2437
##  6       2004       2399
##  7       2005       2877
##  8       2006       3052
##  9       2007       3336
## 10       2008       3120
## 11       2009       3594
## 12       2010       3103
## 13       2011       1927
## 14       2012        775
## 15       2013        313
```

```
LAGOSP_Summary4 <- LAGOSP %>%
  group_by(sampleyear) %>%
```

```
  summarize(tP_Samples = length(tp))

tail(LAGOSP_Summary4, n=15)
```

```
## # A tibble: 15 x 2
##    sampleyear tP_Samples
##         <dbl>      <int>
##  1       1999       4582
##  2       2000       4698
##  3       2001       5949
##  4       2002       5261
##  5       2003       5452
##  6       2004       6156
##  7       2005       6394
##  8       2006       6617
##  9       2007       6451
## 10       2008       6325
## 11       2009       7071
## 12       2010       6913
## 13       2011       5414
## 14       2012       2876
## 15       2013       2624
```

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

> TN: From 2005 to 2010. According to the jitter plot it does not differ much among states. They all seem to have greater concentration of yellow values. Some states such as Wisconsin, Connecticut, and Rhode Island have values more orange, pink, or even purple, meaning that they have more concentration of values from the 1990s, 1980s, and 1970s.
>
> TP: From 2004 to 2010. According to the jitter plot it does differ among states, at least more that the TN plot. There are more "yellow" states (2000s) such as Iowa or Ohio and more "red/orange" states (1990s) like Minnesota or Maine.

Reflection

12. What are 2-3 conclusions or summary points about lake water quality you learned through your analysis?
    1. The trophic state can be calculated using chlorophyll a concentration, Secchi disk transparency, and Total phosphorus (TP).
    2. The results obtained by each one of the variables can be different. chlorophyll a concentration would probably be the most accurate, but Secchi disk is an affordable measurement that can be used as a easy first estimation of the trophic state.
    3. The number of measurements and values of TP and TN can vary considerably between states
13. What data, visualizations, and/or models supported your conclusions from 12?

    For 1. and 2. Calculation of trophic state using the three variables. For 3. all the jitter and violin plots.

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

    It did. I like how hands-on makes you have doubts that you have to do research by yourself to solve them, compared to a theory-based lesson were the professor gives you all the answers in class.

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

    I didn't have much previous knowledge about the topic so I didn't know what to expect. The LAGOS data looks amazing.