Assignment 5: Water Quality in Lakes

Student Name

OVERVIEW

This exercise accompanies the lessons in Water Data Analytics 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, check your PDF against the key and then submit your assignment completion survey at https://forms.gle/fSe18vMhgzcjUKM39

Having trouble? See the assignment's answer key if you need a hint. Please try to complete the assignment without the key as much as possible - this is where the learning happens!

Target due date: 2022-02-22

Setup

1. Verify your working directory is set to the R project file. Load the tidyverse, lubridate, and LAGOSNE packages. Set your ggplot theme (can be theme_classic or something else)
2. Load the LAGOSdata database and the trophic state index csv file we created in class.

```
getwd()
```

```
## [1] "/Users/katerisalk/Box Sync/Courses/Water Data Analytics"
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

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

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(LAGOSNE)

theme_set(theme_classic())

LAGOSdata <- lagosne_load()
```

```
## Warning in (function (version = NULL, fpath = NA) : LAGOSNE version unspecified,
## loading version: 1.087.3
```

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

Trophic State Index

3. 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 <- LAGOStrophic %>%
  mutate(trophic.class.secchi = case_when(TSI.secchi < 40 ~ "Oligotrophic",
                                          TSI.secchi >= 40 & TSI.secchi < 50 ~ "Mesotrophic",
                                          TSI.secchi >= 50 & TSI.secchi< 70 ~ "Eutrophic",
                                          TSI.secchi >= 70 ~ "Hypereutrophic"),
         trophic.class.tp = case_when(TSI.tp < 40 ~ "Oligotrophic",
                                      TSI.tp >= 40 & TSI.tp < 50 ~ "Mesotrophic",
                                      TSI.tp >= 50 & TSI.tp< 70 ~ "Eutrophic",
                                      TSI.tp >= 70 ~ "Hypereutrophic"))
```

4. How many observations fall into the four trophic state categories for the three metrics (trophic.class, trophic.class.secchi, trophic.class.tp)? Hint: `n()` function.

```
LAGOStrophic %>%
  group_by(trophic.class) %>%
  summarise(count = n())
```

```
## # A tibble: 4 x 2
##   trophic.class  count
##   <chr>          <int>
## 1 Eutrophic      37457
## 2 Hypereutrophic 13234
## 3 Mesotrophic    13964
## 4 Oligotrophic    2762
```

```
LAGOStrophic %>%
  group_by(trophic.class.secchi) %>%
  summarise(count = n())
```

```
## # A tibble: 4 x 2
##   trophic.class.secchi count
##   <chr>                <int>
## 1 Eutrophic            25793
## 2 Hypereutrophic        4721
## 3 Mesotrophic          22344
## 4 Oligotrophic         14559
```

```
LAGOStrophic %>%
  group_by(trophic.class.tp) %>%
  summarise(count = n())
```

```
## # A tibble: 4 x 2
##   trophic.class.tp count
##   <chr>            <int>
## 1 Eutrophic        22419
## 2 Hypereutrophic    6407
## 3 Mesotrophic      20607
## 4 Oligotrophic     17984
```

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

```
LAGOStrophic %>%
  group_by(trophic.class) %>%
  summarise(count = n()) %>%
  mutate(prop = count/sum(count))
```

```
## # A tibble: 4 x 3
##   trophic.class  count   prop
##   <chr>          <int>  <dbl>
## 1 Eutrophic      37457 0.556
## 2 Hypereutrophic 13234 0.196
## 3 Mesotrophic    13964 0.207
## 4 Oligotrophic    2762 0.0410
```

```
LAGOStrophic %>%
  group_by(trophic.class.secchi) %>%
  summarise(count = n()) %>%
  mutate(prop = count/sum(count))
```

```
## # A tibble: 4 x 3
##   trophic.class.secchi count   prop
##   <chr>                <int>  <dbl>
## 1 Eutrophic            25793 0.383
## 2 Hypereutrophic        4721 0.0700
## 3 Mesotrophic          22344 0.331
## 4 Oligotrophic         14559 0.216
```

```
LAGOStrophic %>%
  group_by(trophic.class.tp) %>%
  summarise(count = n()) %>%
  mutate(prop = count/sum(count))
```

```
## # A tibble: 4 x 3
##   trophic.class.tp count   prop
##   <chr>            <int>  <dbl>
## 1 Eutrophic        22419 0.333
## 2 Hypereutrophic    6407 0.0950
## 3 Mesotrophic      20607 0.306
## 4 Oligotrophic     17984 0.267
```

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

> TSI as computed from chlorophyll is the most conservative of the metrics, with 56% of samples assigned as eutrophic and 20% of samples assigned as hypereutrophic. Secchi depth and TP assigned lake samples as eutrophic and hypereutrophic 45% and 43% of the time. This outcome may be driven by chlorophyll making up the majority of reduced clarity and primary productivity being limited by P.

Nutrient Concentrations

6. 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. Filter the data frame for May-September. Call this data frame LAGOSNandP.

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



LAGOSNandP <- LAGOSstate %>%
  left_join(., LAGOSlocus) %>%
  left_join(., LAGOSnutrient) %>%
  select(lagoslakeid, sampledate, tn, tp, state, state_name) %>%
  mutate(sampleyear = year(sampledate),
         samplemonth = month(sampledate)) %>%
  filter(samplemonth %in% c(5:9))
```

```
## Joining, by = "state_zoneid"
```

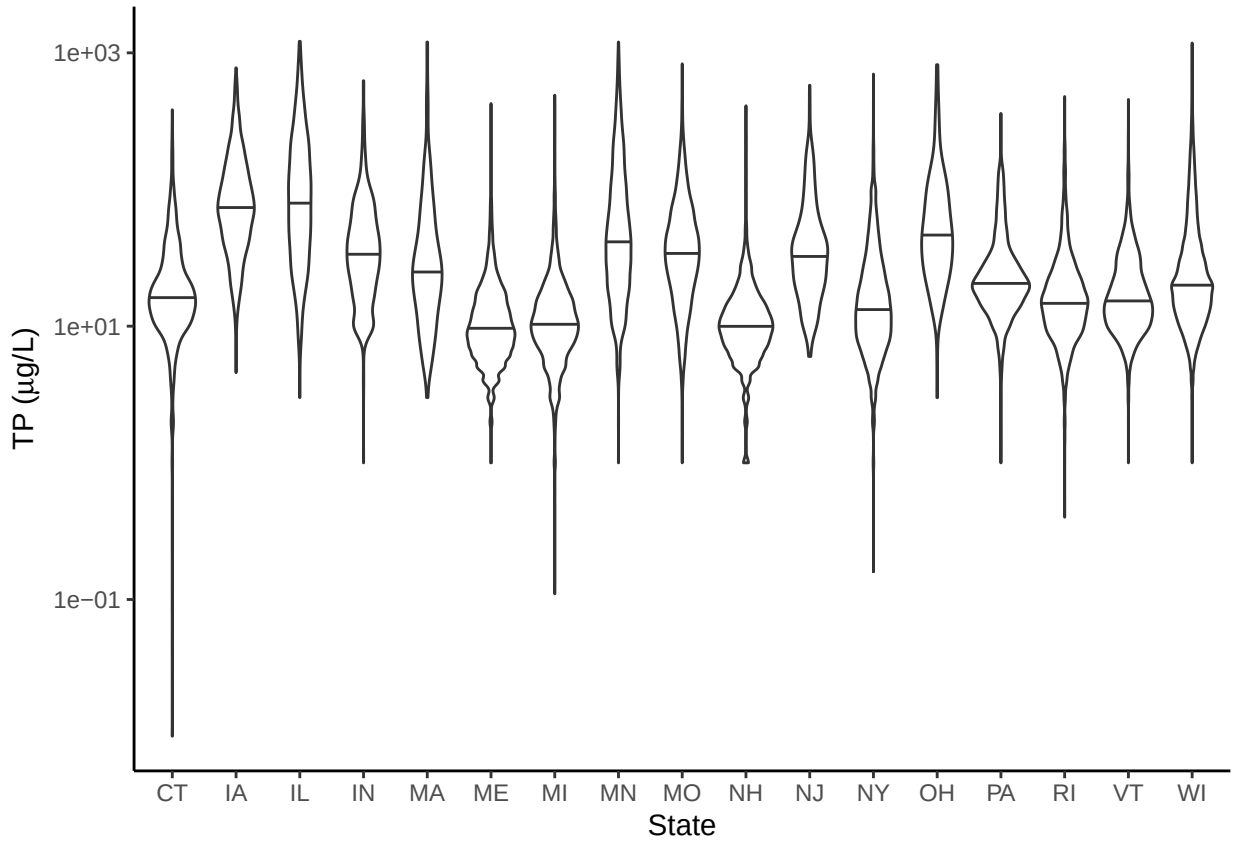
```
## Joining, by = "lagoslakeid"
```

7. Create two violin plots comparing TN and TP concentrations across states. Include a 50th percentile line inside the violins. Create a logged y axis and relabel axes.

```
ggplot(LAGOSNandP, aes(x = state, y = tp)) +
  geom_violin(draw_quantiles = 0.50) +
  scale_y_log10() +
  labs(x = "State", y = expression("TP ("*mu*"g/L)"))
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

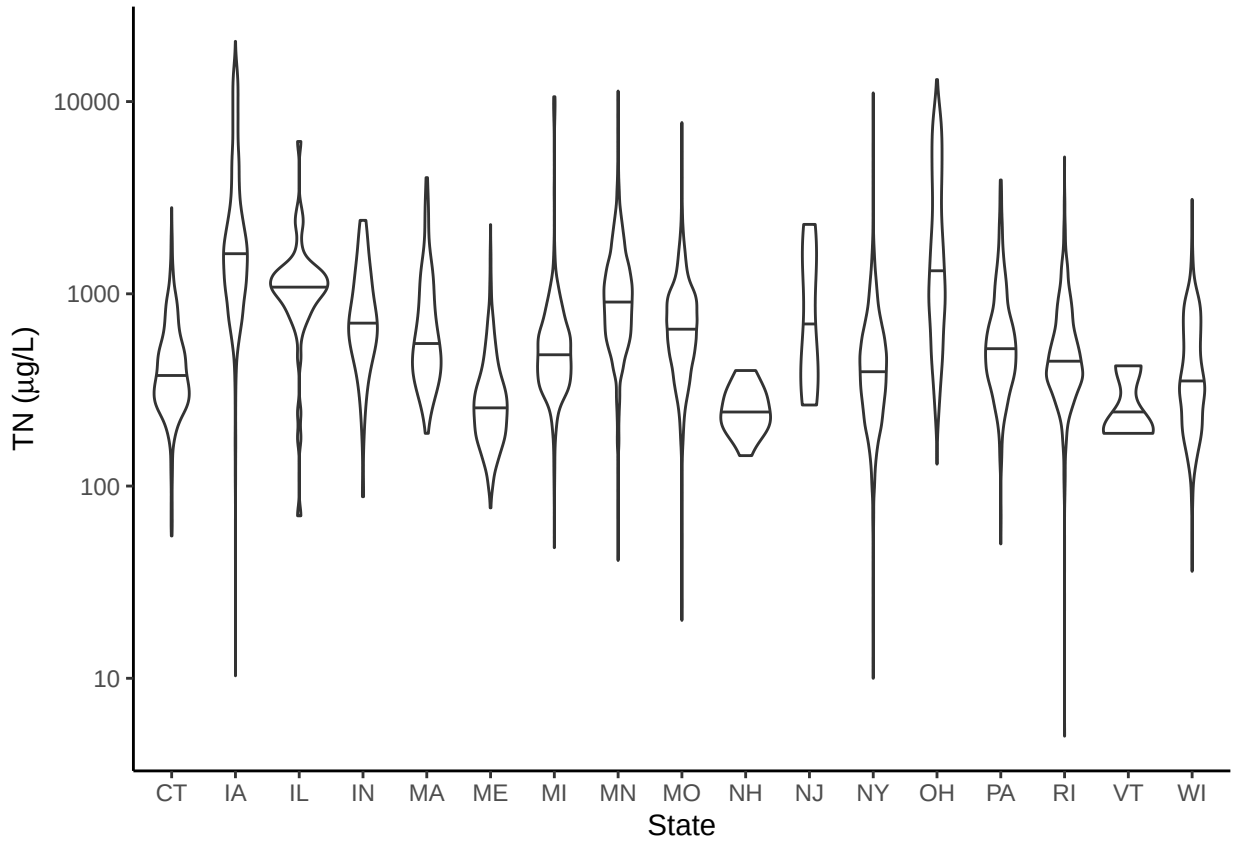
```
## Warning: Removed 618390 rows containing non-finite values (stat_ydensity).
```




```
ggplot(LAGOSNandP, aes(x = state, y = tn)) +
  geom_violin(draw_quantiles = 0.50) +
  scale_y_log10() +
  labs(x = "State", y = expression("TN ("*mu*"g/L)"))
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 697418 rows containing non-finite values (stat_ydensity).
```



Which states have the highest and lowest median concentrations?

TN: highest - IA, OH lowest - CT, ME, NH, VT

TP: highest - IA, IL, OH lowest - CT, ME, MI, NH

Which states have the largest and smallest concentration ranges?

TN: largest - IA, NY, RI smallest - NH, VT

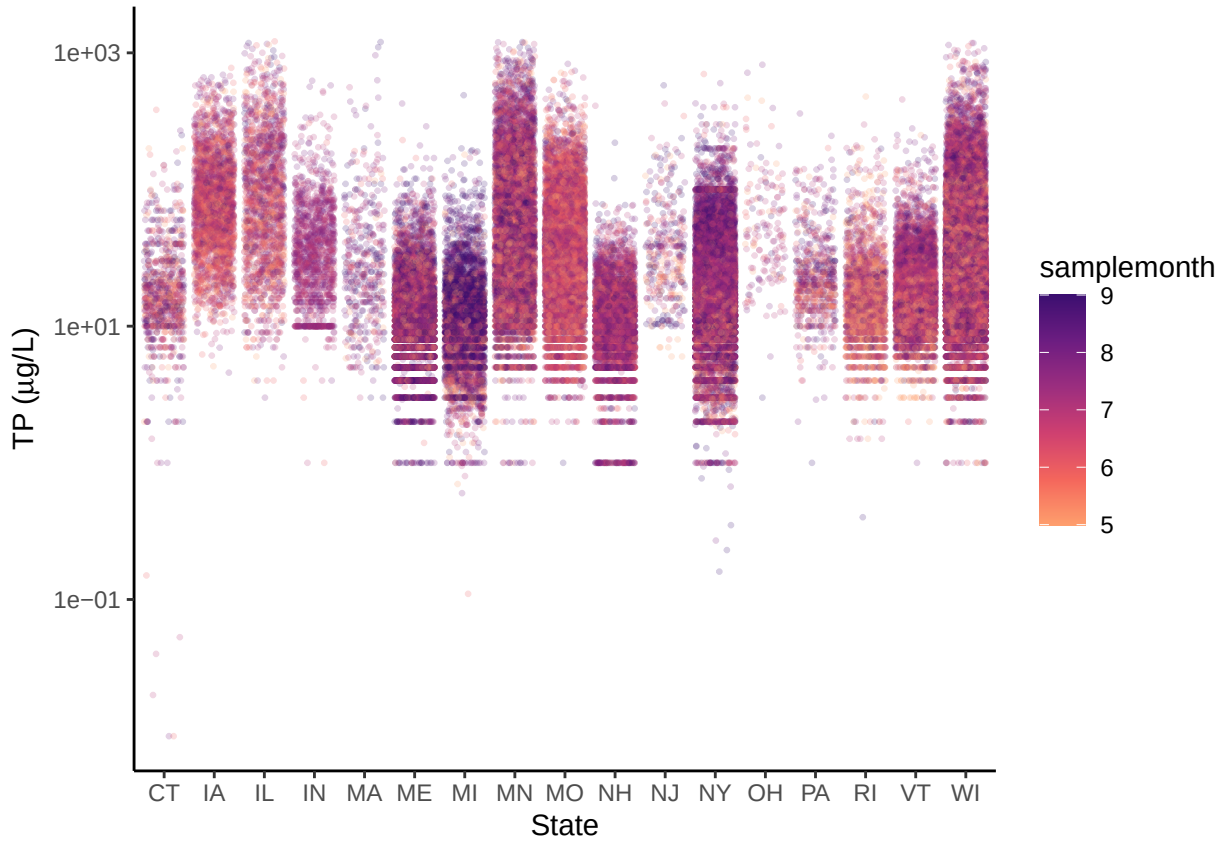
TP: largest - CT smallest - IA, NJ

8. 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.

```
ggplot(LAGOSNandP, aes(x = state, y = tp, color = samplemonth)) +
  geom_jitter(alpha = 0.2, size = 0.5) +
  scale_y_log10() +
  labs(x = "State", y = expression("TP ("*mu*"g/L)")) +
  scale_color_viridis_c(option = "magma", begin = 0.2, end = 0.8, direction = -1)
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

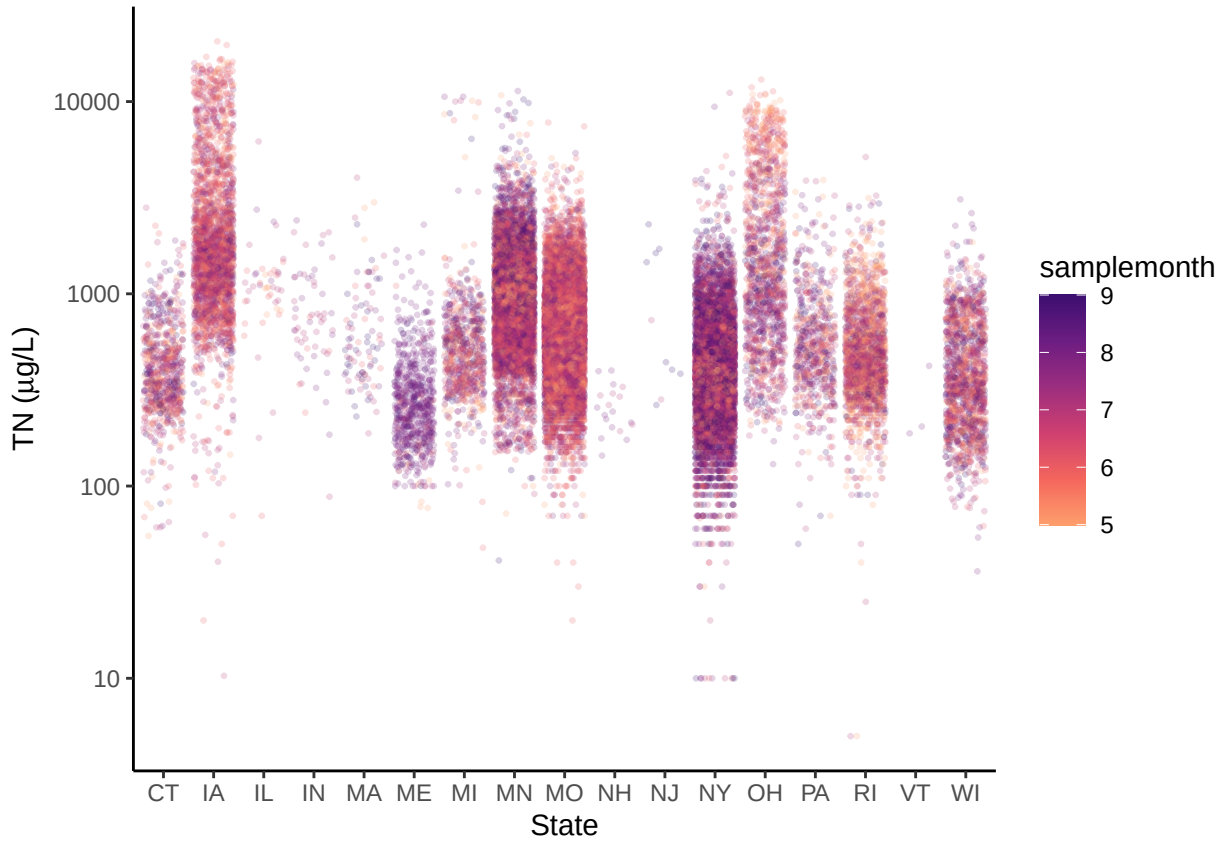
```
## Warning: Removed 618390 rows containing missing values (geom_point).
```




```
ggplot(LAGOSNandP, aes(x = state, y = tn, color = samplemonth)) +
  geom_jitter(alpha = 0.2, size = 0.5) +
  scale_y_log10() +
  labs(x = "State", y = expression("TN ("*mu*"g/L)")) +
  scale_color_viridis_c(option = "magma", begin = 0.2, end = 0.8, direction = -1)
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 697418 rows containing missing values (geom_point).
```



Which states have the fewest samples? How might this have impacted total ranges from #7?

TN: IL, IN, MA, NH, NJ, VT

TP: MA, NJ, OH

Some states with small sample sizes have smaller ranges in TN and TP.