Assignment 5: Water Quality in Lakes

Gabi Richichi

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] "/Users/gabriellerichichi/Documents/5th year @ Duke/Stats/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_classic())

LAGOSdata <- lagosne_load()
```

```
## Warning in `_f`(version = version, fpath = fpath): LAGOSNE version
## unspecified, loading version: 1.087.3
```

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

#The following is code from Lesson 9: Lakes_WaterQuality
LAGOSlocus <- LAGOSdata$locus
LAGOSstate <- LAGOSdata$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")

LAGOSlocations <-
  within(LAGOSlocations,
         state <- factor(state, levels = names(sort(table(state), decreasing=TRUE))))

LAGOStrophic <-
  left_join(LAGOSnutrient, LAGOSlocations, by = "lagoslakeid") %>%
  select(lagoslakeid, sampledate, chla, tp, secchi,
         gnis_name, lake_area_ha, state, state_name) %>%
  mutate(sampleyear = year(sampledate),
         samplemonth = month(sampledate),
         season = as.factor(quarter(sampledate, fiscal_start = 12))) %>%
  drop_na(chla:secchi)
```

```
## Warning: Column `lagoslakeid` joining factors with different levels,
## coercing to character vector
```

```
levels(LAGOStrophic$season) <- c("Winter", "Spring", "Summer", "Fall")

LAGOStrophic <-
  mutate(LAGOStrophic,
         TSI.chl = round(10*(6 - (2.04 - 0.68*log(chla)/log(2)))),
         TSI.secchi = round(10*(6 - (log(secchi)/log(2)))),
         TSI.tp = round(10*(6 - (log(48/tp)/log(2)))),
         trophic.class.chl =
            ifelse(TSI.chl < 40, "Oligotrophic",
                   ifelse(TSI.chl < 50, "Mesotrophic",
                          ifelse(TSI.chl < 70, "Eutrophic", "Hypereutrophic"))))

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

#write.csv(LAGOStrophic, "./Data/LAGOStrophic.csv")#, row.names = FALSE)
```

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.

```
FrameTotal <- LAGOStrophic %>%
  select(trophic.class)
FrameSecchi <- LAGOStrophic %>%
  select(trophic.class.secchi)
FrameTp <- LAGOStrophic %>%
  select(trophic.class.tp)

ObservationsTotal <- count(FrameTotal)
ObservationsSecchi <- count(FrameSecchi)
ObservationsTp <- count(FrameTp)
```

There are 74,951 observations in the LAGOStrophic data frame.

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)?

```
#FrameTotal$trophic.class[1]

#ClassEu <- data.frame(values=numeric())
#ClassEuLoop <- #FrameTotal %<%
#  for(i in FrameTotal$trophic.class){
#  if(i == "Eutrophic"){
    #ClassEu$values[1] = i
#    ClassEu <- rbind(ClassEu, FrameTotal$trophic.class[i])
#  }
#}

ClassEu <- sum(FrameTotal$trophic.class == "Eutrophic")
ClassEu
```

```
## [1] 41861
```

```
ClassHyp <- sum(FrameTotal$trophic.class == "Hypereutrophic")
ClassHyp
```

```
## [1] 14379
```

```
ClassEu_Hyp = ClassEu + ClassHyp
ClassEu_Hyp
```

```
## [1] 56240
```

```
SecchiEu <- sum(FrameSecchi$trophic.class.secchi == "Eutrophic")
SecchiEu
```

```
## [1] 28659
```

```
SecchiHyp <- sum(FrameSecchi$trophic.class.secchi == "Hypereutrophic")
SecchiHyp
```

```
## [1] 5099
```

```
SecchiEu_Hyp = SecchiEu + SecchiHyp
SecchiEu_Hyp
```

```
## [1] 33758
```

```
TpEu <- sum(FrameTp$trophic.class.tp == "Eutrophic")
TpEu
```

```
## [1] 24839
```

```
TpHyp <- sum(FrameTp$trophic.class.tp == "Hypereutrophic")
TpHyp
```

```
## [1] 7228
```

```
TpEu_Hyp = TpEu + TpHyp
TpEu_Hyp
```

```
## [1] 32067
```

```
ClassProp <- ClassEu_Hyp/ObservationsTotal
ClassProp
```

```
##           n
## 1 0.7503569
```

```
SecchiProp <- SecchiEu_Hyp/ObservationsTotal
SecchiProp
```

```
##           n
## 1 0.4504009
```

```
TpProp <- TpEu_Hyp/ObservationsTotal
TpProp
```

```
##           n
## 1 0.4278395
```

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

> The total phosphorus method is the most conservative in its designation of eutrophic conditions. The proportion of measurements considered eutrophic or hypereutrophic in the total phosphorus column is 42.7%, whereas the proportion of measurements considered eutrophic or hypertrophic based on secchi and chlorophyll measurements are 45.0% and 75.0%, respectively. This might be because there are other nutrients that account for eutrophication, such as nitrogen, that measurements for phosphorus do not take into account.

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.

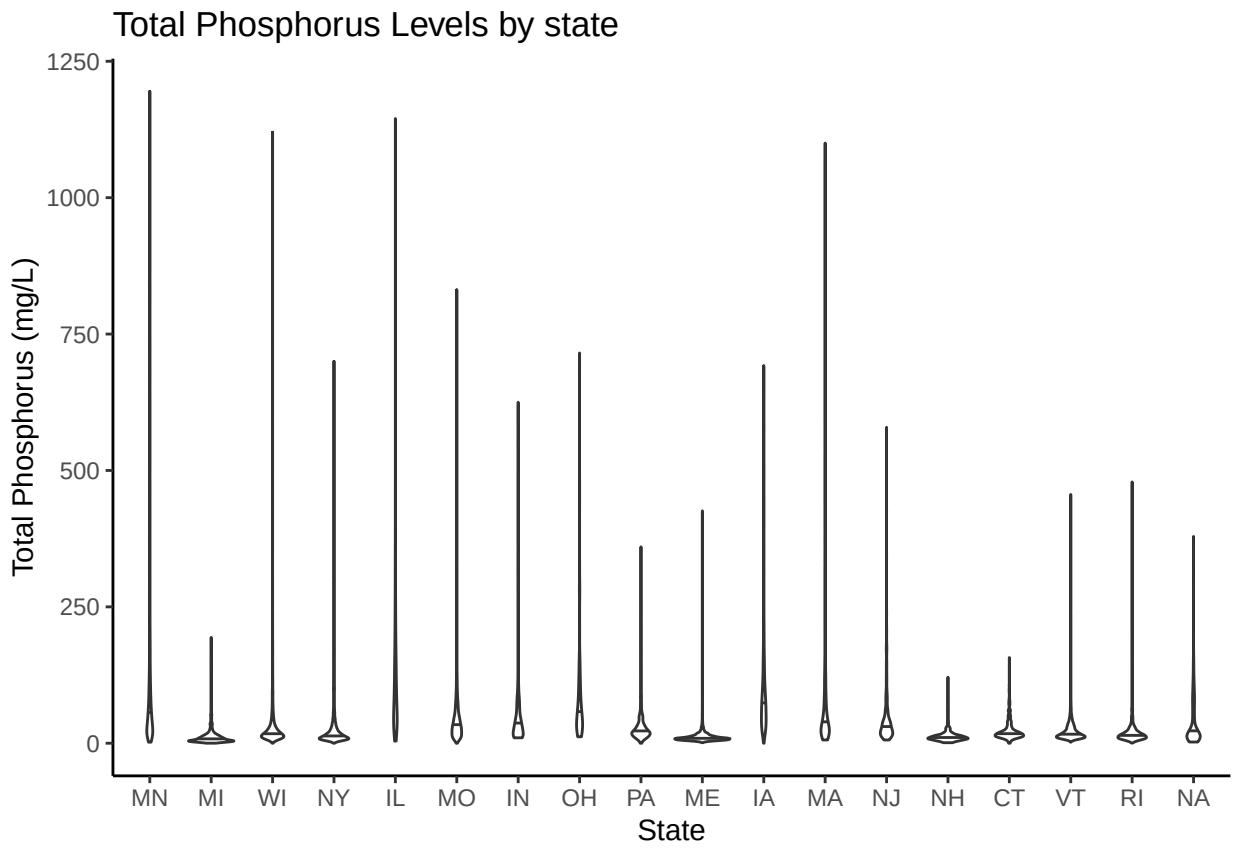
```
LAGOSNandP <- LAGOStrophic %>%
  select(lagoslakeid, sampledate, tp, state, state_name) %>%
  mutate(sampleyear = year(sampledate),
         samplemonth = month(sampledate),
         season = as.factor(quarter(sampledate, fiscal_start = 12)))
#NOTE: left off tn column, because had all "na" values
```

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

```
statetpviolin <- ggplot(LAGOSNandP, aes(x = state, y = tp)) +
  geom_violin(draw_quantiles = 0.50) +
  ggtitle("Total Phosphorus Levels by state") +
  xlab("State") +
  ylab("Total Phosphorus (mg/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

## 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
```




```
#NOTE: did not create a tn plot, because the tn column had all "na" values
```

Which states have the highest and lowest median concentrations?

TN: Did not use tn data, because entire column had “na” values.

TP: Iowa seems to have the highest median concentration of total phosphorus, and Maine seems to have the lowest.

Which states have the highest and lowest concentration ranges?

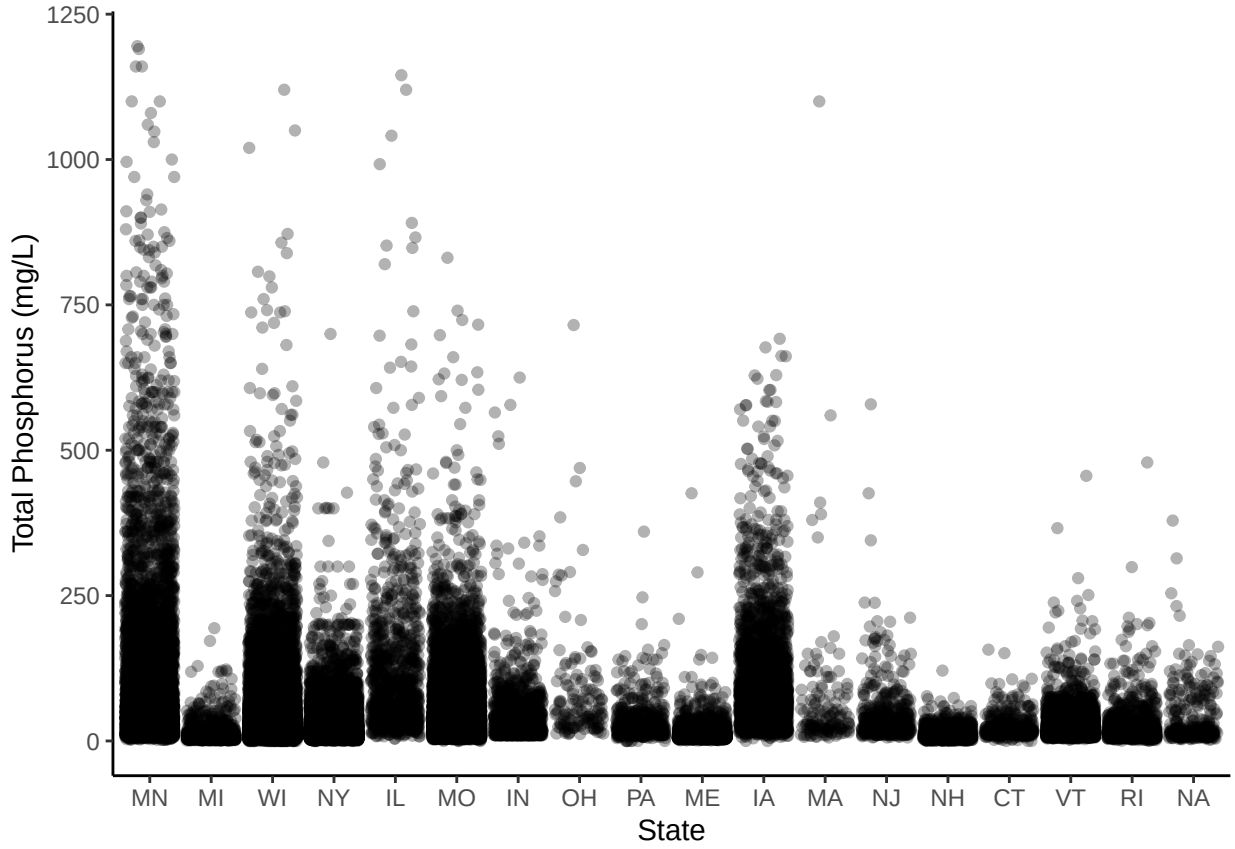
TN: Did not use tn data, because entire column had “na” values.

TP: Minnesota seems to have the highest concentration range, and New Hampshire seems to have the lowest.

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.

```
statetpjitter <- ggplot(LAGOSNandP, aes(x = state, y = tp), color = samplemonth) + #color = samplemonth
 # geom_rect(xmin = -1, xmax = 19, ymin = 0, ymax = 40,
#            fill = "gray90", color = "gray90") +
#  geom_rect(xmin = -1, xmax = 19, ymin = 40, ymax = 50,
#            fill = "gray80", color = "gray80") +
#  geom_rect(xmin = -1, xmax = 19, ymin = 50, ymax = 70,
#            fill = "gray70", color = "gray70") +
#  geom_rect(xmin = -1, xmax = 19, ymin = 70, ymax = 100,
```

```
#            fill = "gray60", color = "gray60") +
  geom_jitter(alpha = 0.3) +
  labs(x = "State", y = "Total Phosphorus (mg/L)")
  #scale_y_continuous() +#limits = c(0, 100)) +
 # theme(legend.position = "top") +
#  scale_color_viridis_d(option = "magma")
print(statetpjitter)
```




```
#NOTE: did not create a tn plot, because the tn column had all "na" values
```

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

TN: Did not use tn data, because entire column had "na" values.

TP: It appears that many of the midwestern states have the most samples. This may have increased their range values from #9.

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

TN:

TP:

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.

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

TN:

TP:

Reflection

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

    I have learned: 1) That chlorophyll levels are a more liberal indication of lake eutrophication, while total phosphorus concentrations are a more conservative indication of eutrophication. 2) That there is considerable variety in the nutrient levels in lakes depending upon the state in which the measurements are being taken. 3) That nutrient levels vary depending on the season.

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

    The violin plot allowed me to visualize the ranges in nutrient levels quite well. The jitter plot allowed me to visualize densities of measurements and time-based / seasonal / monthly changes.

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

    Yes, I believe it did. Coding requires a lot of thinking and problem-solving, which made me more engaged with this material.

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

    It is expected that as nutrient levels increase, so does eutrophication. It is also expected that nutrient levels would vary state-by-state.