Assignment 5: Water Quality in Lakes

Ethan Ready

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 (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "A05_Salk.pdf") 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/ethanready/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
```

```r
library(LAGOSNE)

theme_set(theme_classic())

filename<-file.path(path.expand("~"),"Hydrologic_Data_Analysis", "Data", "Raw", "LAGOSdata.rda")
LAGOS.data<-load(filename)

filename<-file.path(path.expand("~"),"Hydrologic_Data_Analysis", "Data", "LAGOStrophic.csv")
LAGOS.trophic<-read_csv(filename)
```

```
## Parsed with column specification:
## cols(
##   lagoslakeid = col_double(),
##   sampledate = col_date(format = ""),
##   chla = col_double(),
##   tp = col_double(),
##   secchi = col_double(),
##   gnis_name = col_character(),
##   lake_area_ha = col_double(),
##   state = col_character(),
##   state_name = col_character(),
##   sampleyear = col_double(),
##   samplemonth = col_double(),
##   season = col_character(),
##   TSI.chl = col_double(),
##   TSI.secchi = col_double(),
##   TSI.tp = col_double(),
##   trophic.class = col_character()
## )
```

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).

```r
LAGOS.trophic<-LAGOS.trophic%>%
  mutate(trophic.class.secchi =
           ifelse(TSI.secchi<40, "Oligotrophic",
                  ifelse(TSI.secchi<50, "Mesotrophic",
                         ifelse(TSI.secchi<70, "Eutrophic",
                                "Hypereutrophic"))))%>%
  mutate(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.

```
count.chl<-count(LAGOS.trophic, trophic.class)
count.secchi<-count(LAGOS.trophic, trophic.class.secchi)
count.tp<-count(LAGOS.trophic, trophic.class.tp)

count.chl
```

```
## # A tibble: 4 x 2
##   trophic.class      n
##   <chr>          <int>
## 1 Eutrophic      41861
## 2 Hypereutrophic 14379
## 3 Mesotrophic    15413
## 4 Oligotrophic    3298
```

```
count.secchi
```

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

```
count.tp
```

```
## # A tibble: 4 x 2
##   trophic.class.tp     n
##   <chr>            <int>
## 1 Eutrophic        24839
## 2 Hypereutrophic    7228
## 3 Mesotrophic      23023
## 4 Oligotrophic     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 of chl hypereutrophic or eutrophic
sum(count.chl$n[1:2])/sum(count.chl$n)
```

```
## [1] 0.7503569
```

```
#hyper/eu proportion for secchi
sum(count.secchi$n[1:2])/sum(count.secchi$n)
```

```
## [1] 0.4504009
```

```r
#hyper/eu proportion for tp
sum(count.tp$n[1:2])/sum(count.tp$n)
```

```
## [1] 0.4278395
```

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

> Chlorophyll content is far more likely to designate a lake as eutrophic or hypereutrophic than either secchi or total phosphorous. Secchi is slightly more likely than phosphorous to designate a lake as hypereutrophic or eutrophic. This could be due to the formulas for calculating TSI. Or, it could be because chlorophyll could become elevated from factors other than phosphorous, and elevated chlorphyll might not always cloud the water. In those cases, elevated chlorophyll wouldn't show up on the other two measurements.

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.

```r
LAGOSNandP<-
  left_join(LAGOSdata$locus, LAGOSdata$state, by="state_zoneid")%>%
  left_join(LAGOSdata$epi_nutr, by="lagoslakeid")%>%
  select(lagoslakeid, sampledate, tn, tp, state, state_name)%>%
  mutate(sampleyear = year(sampledate), samplemonth = month(sampledate))
```

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

```r
ggplot(LAGOSNandP)+
  geom_violin(aes(x=state, y=tn), draw_quantiles = 0.50)+
  labs(y= "Total Nitrogen")
```

```
## Warning: Removed 900834 rows containing non-finite values (stat_ydensity).

## 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
```

```r
ggplot(LAGOSNandP)+
  geom_violin(aes(x=state, y=tp), draw_quantiles = 0.50)+
  labs(y="Total Phosphorous")
```

```
## Warning: Removed 799469 rows containing non-finite values (stat_ydensity).

## Warning: collapsing to unique 'x' values

## Warning: collapsing to unique 'x' values

## Warning: collapsing to unique 'x' values

## Warning: collapsing to unique 'x' values

## Warning: collapsing to unique 'x' values

## Warning: collapsing to unique 'x' values
```

```
## Warning: collapsing to unique 'x' values

## Warning: collapsing to unique 'x' values

## Warning: collapsing to unique 'x' values

## Warning: collapsing to unique 'x' values

## Warning: collapsing to unique 'x' values

## Warning: collapsing to unique 'x' values

## Warning: collapsing to unique 'x' values
```

Which states have the highest and lowest median concentrations?

TN: Highest: IA, IL Lowest: ME, NH, VT

TP: Highest: IL, IA Lowest: ME, MI, NH, VT

Which states have the highest and lowest concentration ranges?

TN: Highest: IA Lowest: NH, VT

TP: Highest: MN Lowest: PA, ME, CT

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.

```
jitter.month.tn<-ggplot(LAGOSNandP, aes(x=state, y=tn, color=as.factor(samplemonth)))+
  geom_jitter()+
  labs(y="Total Nitrogen")+
  scale_color_viridis_d(option = "magma")

jitter.month.tp<-ggplot(LAGOSNandP, aes(x=state, y=tp, color=as.factor(samplemonth)))+
  geom_jitter()+
  labs(y="Total Phosphorous")+
  scale_color_viridis_d(option = "magma")

jitter.month.tn
```

```
## Warning: Removed 900834 rows containing missing values (geom_point).
```

```
jitter.month.tp
```

```
## Warning: Removed 799469 rows containing missing values (geom_point).
```

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

> TN: IA, MN, MO, WI have the most sample while VT and NH have very few. This corresponds to the higher range for IA and lower ranges for VT and NH.
>
> TP: MN, WI are most extensive. MN had the highest range so this makes sense.

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

> TN: Summer months are sampled most extensively. Wisconsin samples the more during the winter.
>
> TP: Summer months are sampled most extensively. Wisconsin samples the more during the winter.

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.

```
jitter.year.tn<-ggplot(LAGOSNandP, aes(x=state, y=tn, color=as.factor(sampleyear)))+
  geom_jitter()+
  labs(y="Total Nitrogen")+
  scale_color_viridis_d(option = "magma")

jitter.year.tp<-ggplot(LAGOSNandP, aes(x=state, y=tp, color=as.factor(sampleyear)))+
```

```
  geom_jitter()+
  labs(y="Total Phosphorous")+
  scale_color_viridis_d(option = "magma")

jitter.year.tn
```

```
## Warning: Removed 900834 rows containing missing values (geom_point).
```

```
jitter.year.tp
```

```
## Warning: Removed 799469 rows containing missing values (geom_point).
```

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

> TN: Generally more recent years are sampled more often. Minnesota and Conneticut have some older data.
>
> TP: Generally more recent years are sampled more often. Minnesota has lots of older data.

Reflection

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

    > Lots of lakes are either hypereutrophic or eutrophic. It can be hard to determine the trophic class of a lake because there are different metrics.

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

    > The analysis of secchi, chla, and tp percentage of hypereutrophic/eutrophic.

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

    > Yes it made me more comfortable with code and finding these things myself.

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

    > More lakes are polluted