Assignment 6: Water Quality in Lakes

Lydie Costes

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/swoene3wmGGVUnnm7

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: 2022-03-01

Setup

1. Verify your working directory is set to the R project file. Load the tidyverse, lubridate, and dataRetrieval packages. Set your ggplot theme (can be theme_classic or something else)
2. Use the dataRetrieval package to fetch North Carolina lake TN, TP, and chlorophyll data from the Water Quality Portal. Since we didn't use this method in class, the code is provided for you. Be patient; this query will take some time.

General Data Import from Water Quality Portal Water Quality Portal Web Services Guide

```
# 1. Check directory, load packages, set theme
getwd()
```

```
## [1] "/Users/lydiecostes/Documents/Duke/WaterDataAnalytics/Water_Data_Analytics_2022/Assignments"
```

```
library(lubridate)
```

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

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

```
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 lubridate::as.difftime() masks base::as.difftime()
## x lubridate::date()        masks base::date()
## x dplyr::filter()          masks stats::filter()
## x lubridate::intersect()   masks base::intersect()
## x dplyr::lag()             masks stats::lag()
## x lubridate::setdiff()     masks base::setdiff()
## x lubridate::union()       masks base::union()
```

```
library(dataRetrieval)
theme_set(theme_bw())

# 2. Load data
NCdata <- readWQPdata(statecode = "North Carolina",
                      siteType = "Lake, Reservoir, Impoundment",
                      sampleMedia = c("Water", "water"),
                      CharacteristicName = c("Phosphorus", "Chlorophyll a"))

NCsites <- whatWQPsites(statecode = "North Carolina",
                      siteType = "Lake, Reservoir, Impoundment")
```

Water Quality Portal downloads have the same columns each time, but be aware that data are uploaded to the Water Quality Portal by individual organizations, which may or may not follow the same conventions. Data and metadata quality are not guaranteed! Make sure to carefully explore any data and make conservative quality assurance decisions where information is limited.

General data processing and quality assurance considerations:

- WQP data is acquired in long format. It may be useful to wrangle the dataset into wide format (we will do this today)
- `readWQPdata` does not inherently restrict the variables pulled from WQP. You may specify the desired variables by using, for instance: 'characteristicName = "pH" '
- **ResultMeasureValue** should be numeric, with details on detection limits, qualifiers, etc. provided in other columns. This is not always the case!
- **ResultSampleFractionText** specifies forms of constituents. In some cases, a single **CharacteristicName** will have both "Total" and "Dissolved" forms specified, which should not be combined.
- Some variables have different names but represent the same constituent (e.g., "Total Kjeldahl nitrogen (Organic N & NH3)" and "Kjeldahl nitrogen"). Always refer to the **ResultAnalyticalMethod** columns to verify methods are measuring the same constituent.
- **ActivityDepthHeightMeasure.MeasureValue** provides depth information. This is a crucial column for lake data but less often for river data.

- **ResultCommentText** often has details relating to additional QA.
- **MeasureQualifierCode** Contains information about data flags:
- *U* designates below detection limit (action: set value to 1/2 detection or quantitation limit from **DetectionQuantitationLimitMeasure.MeasureValue**)
- *J* designates above detection limit but below quantitation limit (action: retain value)
- Other codes may designate suspect data or other flags which may be described in detail in **ResultLaboratoryCommentText** or another column

3. Note the wrangling steps outlined here. Then, set ActivityStartDate as date.

```
# 3. Set ActivityStartDate as date
NCdata$ActivityStartDate <- as.Date(NCdata$ActivityStartDate)

# Look into what forms and units phosphorus and chlorophyll have.
#   Phosphorus = Total or Dissolved; mg/l as P
#   Chlorophyll = NA; ug/l
# Note that chlorophyll is always suspended, so the form isn't crucial.
# Note that Phosphorus is in dissolved and total forms. We are interested in total.
# Note that chlorophyll is in ug/L and phosphorus is in mg/L. In this case, we will assume NAs are cons
# Note that we have 227 "U" flags in the dataset. This means the concentration was below the detection
summary(as.factor(NCdata$ResultSampleFractionText[NCdata$CharacteristicName == "Chlorophyll a"]))
```

```
## Filtered, lab         Total          NA's
##            30          8782         15002
```

```
summary(as.factor(NCdata$ResultSampleFractionText[NCdata$CharacteristicName == "Phosphorus"]))
```

```
## Dissolved     Total      NA's
##      1104     20863        27
```

```
summary(as.factor(NCdata$ResultMeasure.MeasureUnitCode[NCdata$CharacteristicName == "Phosphorus"]))
```

```
##      mg/l mg/l as P      NA's
##     15852      4199      1943
```

```
summary(as.factor(NCdata$ResultSampleFractionText[NCdata$CharacteristicName == "Phosphorus"]))
```

```
## Dissolved     Total      NA's
##      1104     20863        27
```

```
summary(as.factor(NCdata$MeasureQualifierCode))
```

```
##     H     U  NA's
##     4   227 45577
```

```
# ResultMeasureValue should be numeric. Note that there are a few non-numeric entries in that column, s
NCdata$ResultMeasureValue <- as.numeric(NCdata$ResultMeasureValue)
```

```
## Warning: NAs introduced by coercion
```

```
# Depth is sometimes reported in feet and sometimes in meters. Make a new column with all depths in m.
# Create a new variable column which notes the form and units of the nutrient.
# Update the ResultMeasureValue column with 1/2 detection limit for U flag rows. Retain other samples a
NCdata_processed <- NCdata %>%
  mutate(Depth_m = case_when(ActivityDepthHeightMeasure.MeasureUnitCode %in% c("feet", "ft") ~
                             ActivityDepthHeightMeasure.MeasureValue * 0.3048,
                             ActivityDepthHeightMeasure.MeasureUnitCode %in% c("meters", "m") ~
                               ActivityDepthHeightMeasure.MeasureValue),
         Variable = case_when(CharacteristicName == "Phosphorus" &
                                ResultSampleFractionText == "Total" ~ "TP_mgL",
                              CharacteristicName == "Phosphorus" &
                                ResultSampleFractionText == "Dissolved" ~ "TDP_mgL",
                              CharacteristicName == "Chlorophyll a" ~ "Chla_ugL"),
         ResultMeasureValue = case_when(MeasureQualifierCode == "U" ~
                                          DetectionQuantitationLimitMeasure.MeasureValue,
                                        TRUE ~ ResultMeasureValue)) %>%
  drop_na(Variable, ResultMeasureValue, Depth_m)
```

4. Then, add a new pipe that does the following:

- Select the columns OrganizationIdentifier, OrganizationFormalName, ActivityStartDate, ActivityConductingOrganizationText, MonitoringLocationIdentifier, Depth, ResultMeasureValue, Variable
- Group by those same variables except ResultMeasureValue, then `summarise` the mean of ResultMeasureValue. This will generate a single row for each location-date-depth sample.
- Pivot your dataset into wide format. Hint: `pivot_wider` by Variable and ResultMeasureValue.
- Add in month and year as new columns. Filter the dataset for just months May-October.
- Filter the dataset for just depths 1 m or shallower.

5. Create a graph visualizing the chlorophyll-TP stressor-response relationship. Add a line of best fit for the linear regression, and adjust axis scales and labels as needed.

```
# 4. Add new pipe
NCdata_processed_wider <- NCdata_processed %>%
  select(OrganizationIdentifier, OrganizationFormalName, ActivityStartDate,
         ActivityConductingOrganizationText, MonitoringLocationIdentifier,
         Depth_m, ResultMeasureValue, Variable) %>%
  group_by(OrganizationIdentifier, OrganizationFormalName, ActivityStartDate,
           ActivityConductingOrganizationText, MonitoringLocationIdentifier,
           Depth_m, Variable) %>%
  summarise(mean = mean(ResultMeasureValue)) %>%
  pivot_wider(names_from = Variable, values_from = mean) %>%
  mutate(month = month(ActivityStartDate),
         year = year(ActivityStartDate)) %>%
  filter(month %in% c(5:9) & Depth_m <= 1)
```

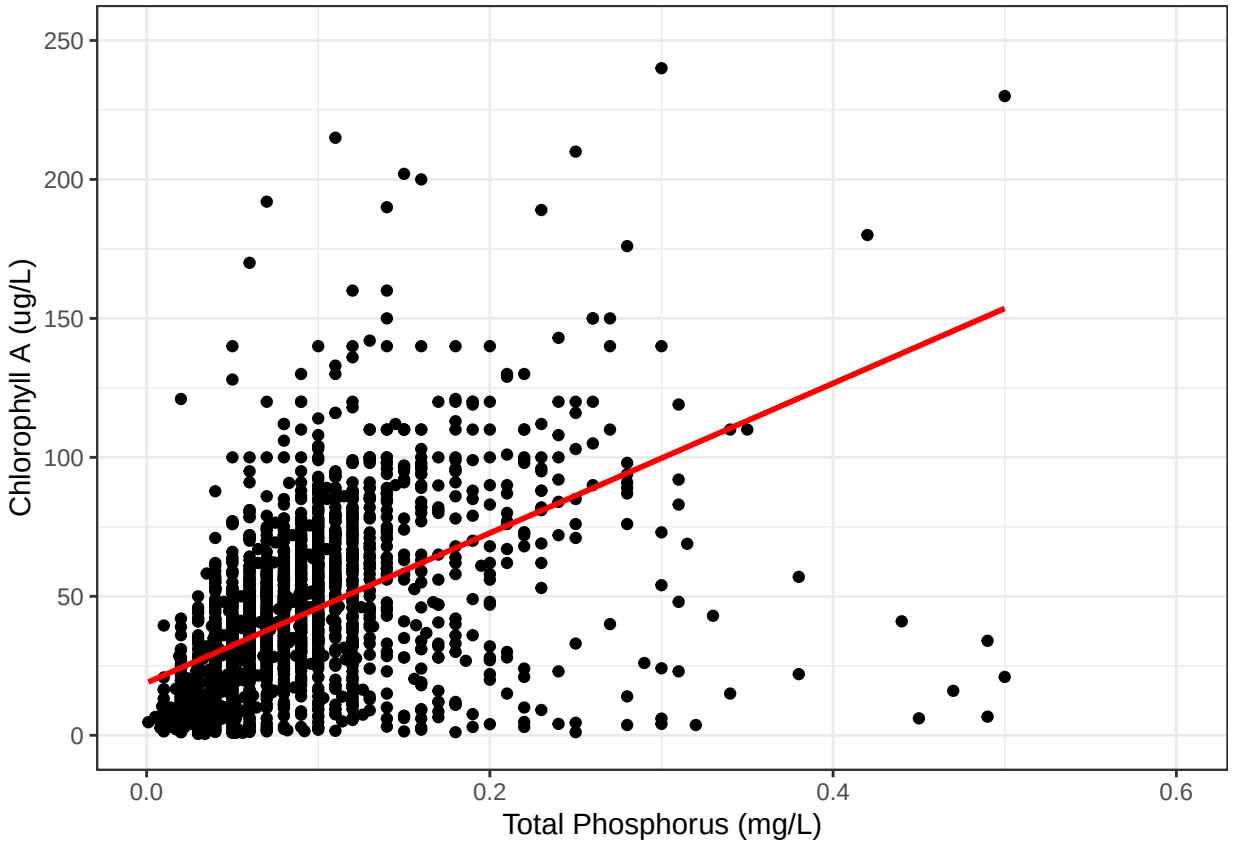
```
## `summarise()` has grouped output by 'OrganizationIdentifier', 'OrganizationFormalName', 'ActivitySta
```

```r
# 5. Create graph
ggplot(NCdata_processed_wider, aes(x=TP_mgL, y=Chla_ugL)) +
  geom_point() +
  geom_smooth(method="lm", color="red", se=FALSE) +
  xlim(NA, 0.6) +
  ylim(NA, 250) +
  labs(x="Total Phosphorus (mg/L)", y="Chlorophyll A (ug/L)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 2139 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 2139 rows containing missing values (geom_point).
```



6. How might you expect this stressor-response relationship to change (or not) across the Mountains, Piedmont, and Coastal Plain regions of the state? How might you anticipate natural lakes might behave differently than reservoirs?

> Because the mountains experience more climatic variability than the coastal plain, I would guess that we might see a flatter regression line. In the warm months, there would be a strong relationship between P and Chl, whereas since the mountains have more temps below freezing, we would expect the Chl levels to drop while the P levels might remain more steady. I'm not sure how to answer the natural lake/reservoir question. Perhaps reservoirs have higher turbidity from increased erosion and sediment load? That could result in reduced algal levels if there isn't enough light penetration to allow for photosynthesis.