Assignment 3: Data Exploration

Caroline Watson

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics (ENV872L) on data exploration.

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Use the lesson as a guide. It contains code that can be modified to complete the assignment.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure to **answer the questions** in this assignment document. Space for your answers is provided in this document and is indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">". You should notice that the answer is highlighted in green by RStudio.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file. You will need to have the correct software installed to do this (see Software Installation Guide) Press the `Knit` button in the RStudio scripting panel. This will save the PDF output in your Assignments folder.
6. After Knitting, please submit the completed exercise (PDF file) to the dropbox in Sakai. Please add your last name into the file name (e.g., "Salk_A02_DataExploration.pdf") prior to submission.

The completed exercise is due on Thursday, 31 January, 2019 before class begins.

1) Set up your R session

Check your working directory, load necessary packages (tidyverse), and upload the North Temperate Lakes long term monitoring dataset for the light, temperature, and oxygen data for three lakes (file name: NTL-LTER_Lake_ChemistryPhysics_Raw.csv). Type your code into the R chunk below.

```
getwd()
```

```
## [1] "/Users/carolinewatson/Documents/Spring 2019/Environmental Data Analytics/Env_Data_Analytics/Ass
```

```
suppressMessages(library(tidyverse))
NTLLTER.Lake.Chem.Physics.data <- read.csv("../Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv")
```

2) Learn about your system

Read about your dataset in the NTL-LTER README file. What are three salient pieces of information you gained from reading this file?

> ANSWER: The data were collected from the North Temperate Lakes Long Term Ecological Research website on studies of Lakes in Wisconsin. Information that is found in the database (Carbon, Nutrients, and Physical and Chemical Limnology). The file also gives an explination on how the file is named so someone can understand what this file contains and the stage the data is in.

3) Obtain basic summaries of your data

Write R commands to display the following information:

1. dimensions of the dataset
2. class of the dataset
3. first 8 rows of the dataset
4. class of the variables lakename, sampledate, depth, and temperature
5. summary of lakename, depth, and temperature

```
#1 - dimensions of the dataset
dim(NTLLTER.Lake.Chem.Physics.data)
```

```
## [1] 38614    11
```

```
# 2 - class of the dataset
class(NTLLTER.Lake.Chem.Physics.data)
```

```
## [1] "data.frame"
```

```
# 3 - first 8 rows of the dataset
head(NTLLTER.Lake.Chem.Physics.data, 8)
```

```
##   lakeid  lakename year4 daynum sampledate depth temperature_C
## 1      L Paul Lake  1984    148    5/27/84  0.00          14.5
## 2      L Paul Lake  1984    148    5/27/84  0.25            NA
## 3      L Paul Lake  1984    148    5/27/84  0.50            NA
## 4      L Paul Lake  1984    148    5/27/84  0.75            NA
## 5      L Paul Lake  1984    148    5/27/84  1.00          14.5
## 6      L Paul Lake  1984    148    5/27/84  1.50            NA
## 7      L Paul Lake  1984    148    5/27/84  2.00          14.2
## 8      L Paul Lake  1984    148    5/27/84  3.00          11.0
##   dissolvedOxygen irradianceWater irradianceDeck comments
## 1             9.5            1750           1620     <NA>
## 2              NA            1550           1620     <NA>
## 3              NA            1150           1620     <NA>
## 4              NA             975           1620     <NA>
## 5             8.8             870           1620     <NA>
## 6              NA             610           1620     <NA>
## 7             8.6             420           1620     <NA>
## 8            11.5             220           1620     <NA>
```

```
# 4
#class of lakename
class(NTLLTER.Lake.Chem.Physics.data$lakename)
```

```
## [1] "factor"
```

```
#class of sampledate
class(NTLLTER.Lake.Chem.Physics.data$sampledate)
```

```
## [1] "factor"
```

```
#class of depth
class(NTLLTER.Lake.Chem.Physics.data$depth)
```

```
## [1] "numeric"
```

```
#class of temperature
class(NTLLTER.Lake.Chem.Physics.data$temperature_C)
```

```
## [1] "numeric"
```

```r
# 5
#summary of lakename
summary(NTLLTER.Lake.Chem.Physics.data$lakename)
```

```
## Central Long Lake     Crampton Lake    East Long Lake  Hummingbird Lake 
##               539              1234              3905               430 
##         Paul Lake        Peter Lake      Tuesday Lake         Ward Lake 
##             10325             11288              6107               598 
##    West Long Lake 
##              4188
```

```r
#summary of depth
summary(NTLLTER.Lake.Chem.Physics.data$depth)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00    1.50    4.00    4.39    6.50   20.00
```

```r
#summary of temperature
summary(NTLLTER.Lake.Chem.Physics.data$temperature_C)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    0.30    5.30    9.30   11.81   18.70   34.10    3858
```

Change sampledate to class = date. After doing this, write an R command to display that the class of sammpledate is indeed date. Write another R command to show the first 10 rows of the date column.

```r
#changing sampledate to class = date from class = factor
NTLLTER.Lake.Chem.Physics.data$sampledate <- as.Date(NTLLTER.Lake.Chem.Physics.data$sampledate,
        format = "%m/%d/%y")

#checking the class of sampledate
class(NTLLTER.Lake.Chem.Physics.data$sampledate)
```

```
## [1] "Date"
```

```r
#showing the first 10 rows of the date column
head(NTLLTER.Lake.Chem.Physics.data$sampledate, 10)
```

```
##  [1] "1984-05-27" "1984-05-27" "1984-05-27" "1984-05-27" "1984-05-27"
##  [6] "1984-05-27" "1984-05-27" "1984-05-27" "1984-05-27" "1984-05-27"
```

Question: Do you want to remove NAs from this dataset? Why or why not?

> ANSWER: No, we do not want to remove NAs from this dataset because the NAs most likely signify that measurements (such as temperature and DO) were not recorded at that date/time when other informaiton was recorded. Also, when using R, we are able to plot information to explore our data and the NAs will be left out of the plots.

4) Explore your data graphically

Write R commands to display graphs depicting:

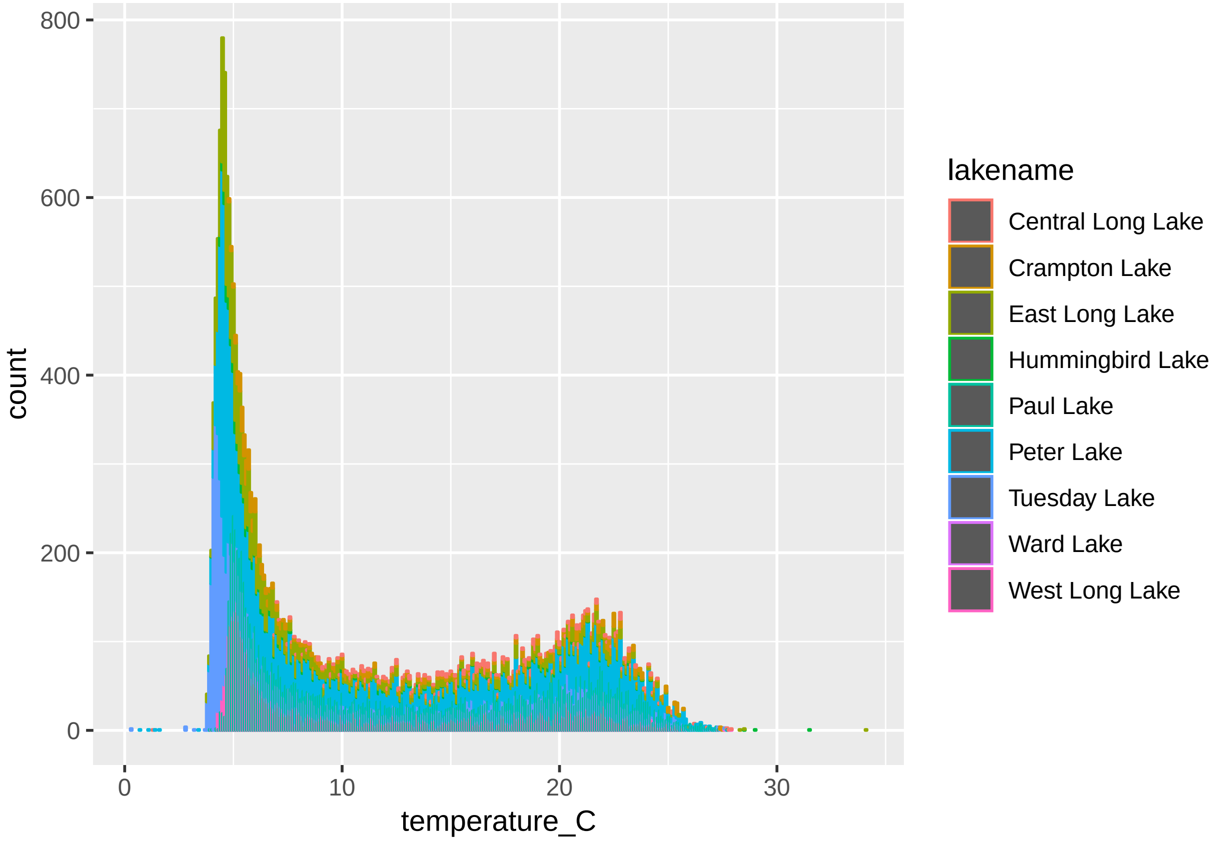
1. Bar chart of temperature counts for each lake
2. Histogram of count distributions of temperature (all temp measurements together)
3. Change histogram from 2 to have a different number or width of bins
4. Frequency polygon of temperature for each lake. Choose different colors for each lake.
5. Boxplot of temperature for each lake
6. Boxplot of temperature based on depth, with depth divided into 0.25 m increments

7. Scatterplot of temperature by depth

```
# 1 - Creating bar chart of temperature counts for each lake
ggplot(NTLLTER.Lake.Chem.Physics.data, aes(x = temperature_C, color = lakename)) + geom_bar()
```

```
## Warning: Removed 3858 rows containing non-finite values (stat_count).
```

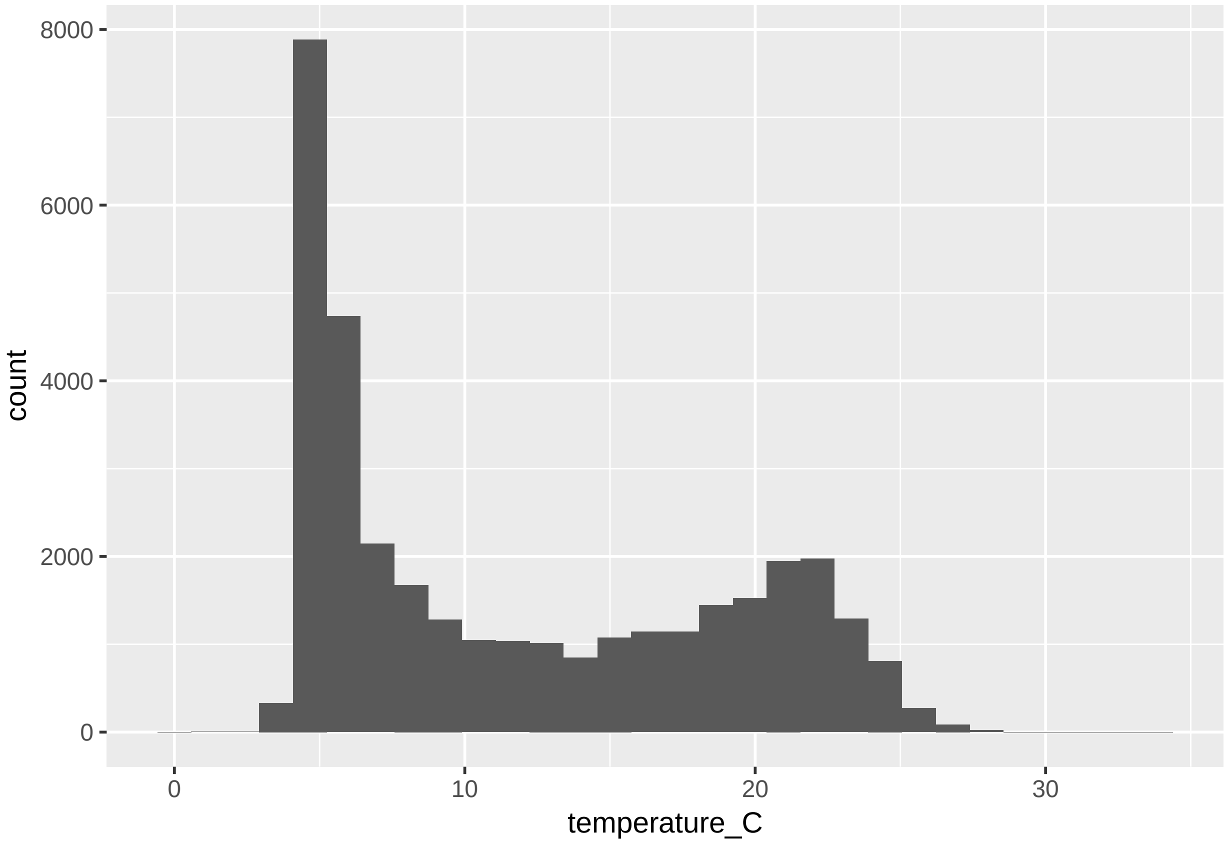
```
## Warning: position_stack requires non-overlapping x intervals
```



```
# 2 #histogram of temperature data
ggplot(NTLLTER.Lake.Chem.Physics.data) + geom_histogram(aes(x = temperature_C))
```

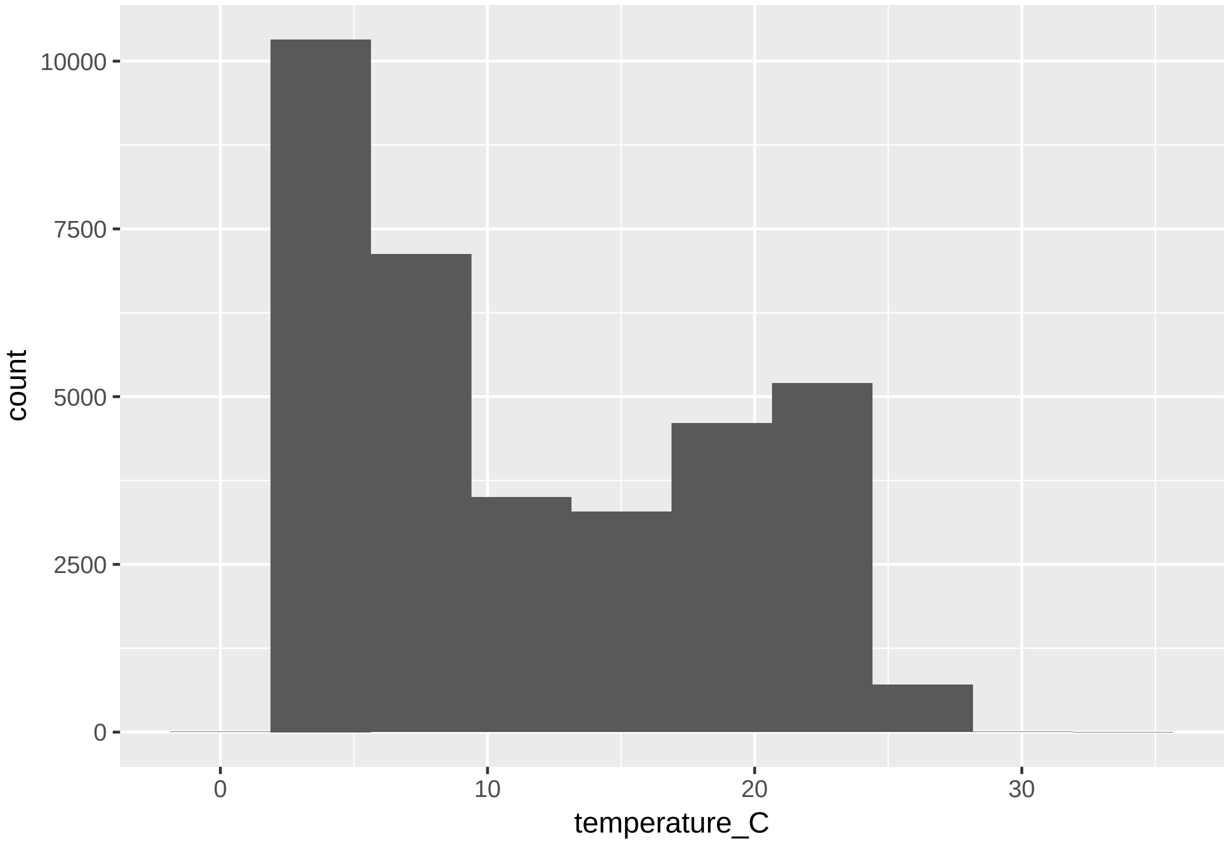
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 3858 rows containing non-finite values (stat_bin).
```



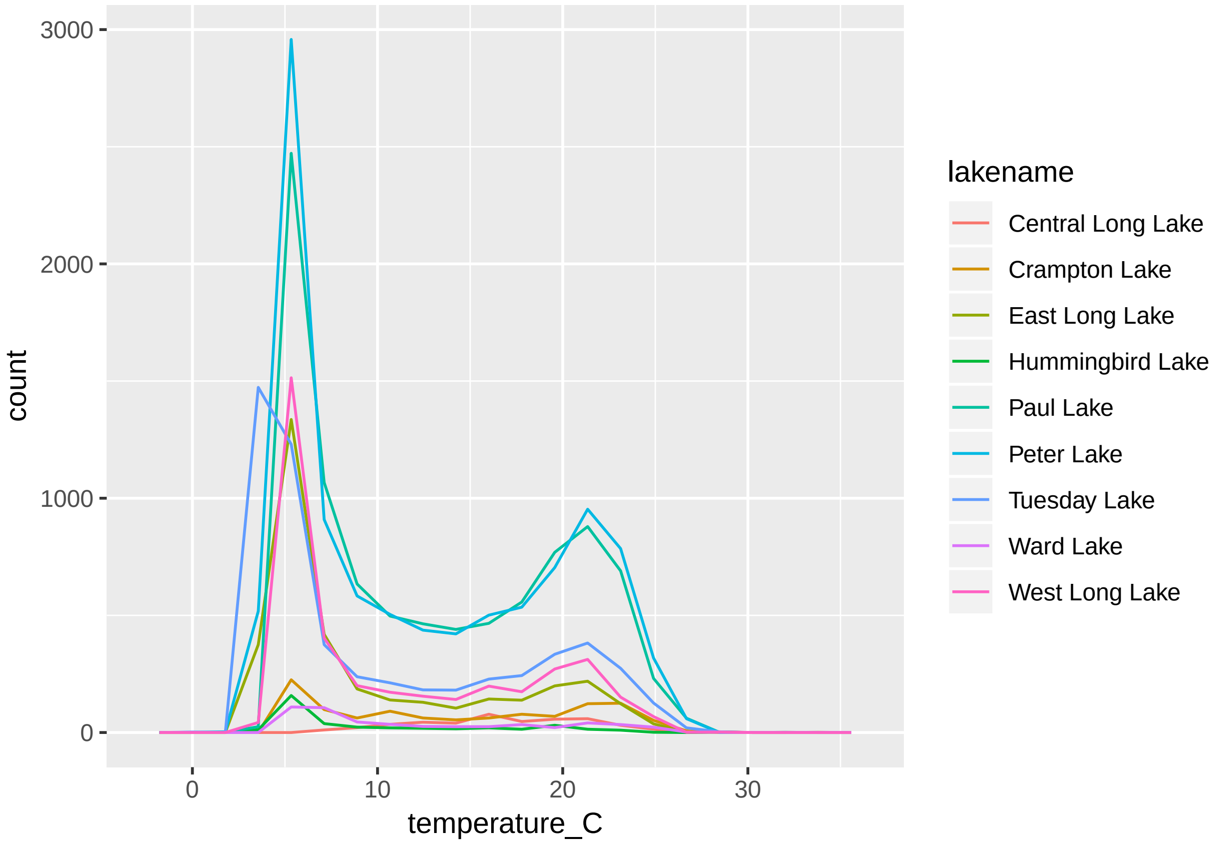
```
# 3 #histogram of temperature data
ggplot(NTLLTER.Lake.Chem.Physics.data) + geom_histogram(aes(x = temperature_C), bins = 10)
```

```
## Warning: Removed 3858 rows containing non-finite values (stat_bin).
```



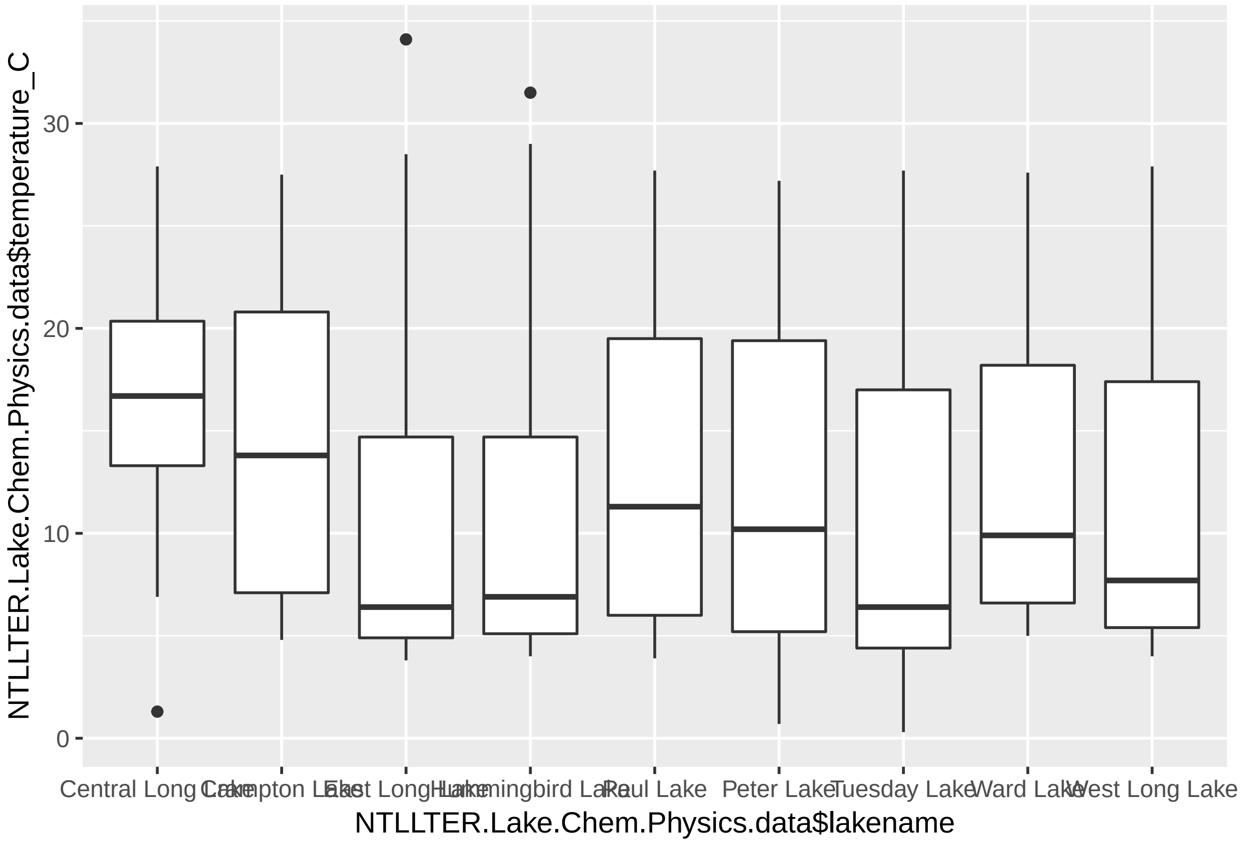
```
# 4 #frequency polygon of temperature for each lake
ggplot(NTLLTER.Lake.Chem.Physics.data) +
  geom_freqpoly(aes(x = temperature_C, color = lakename), bins = 20)
```

```
## Warning: Removed 3858 rows containing non-finite values (stat_bin).
```



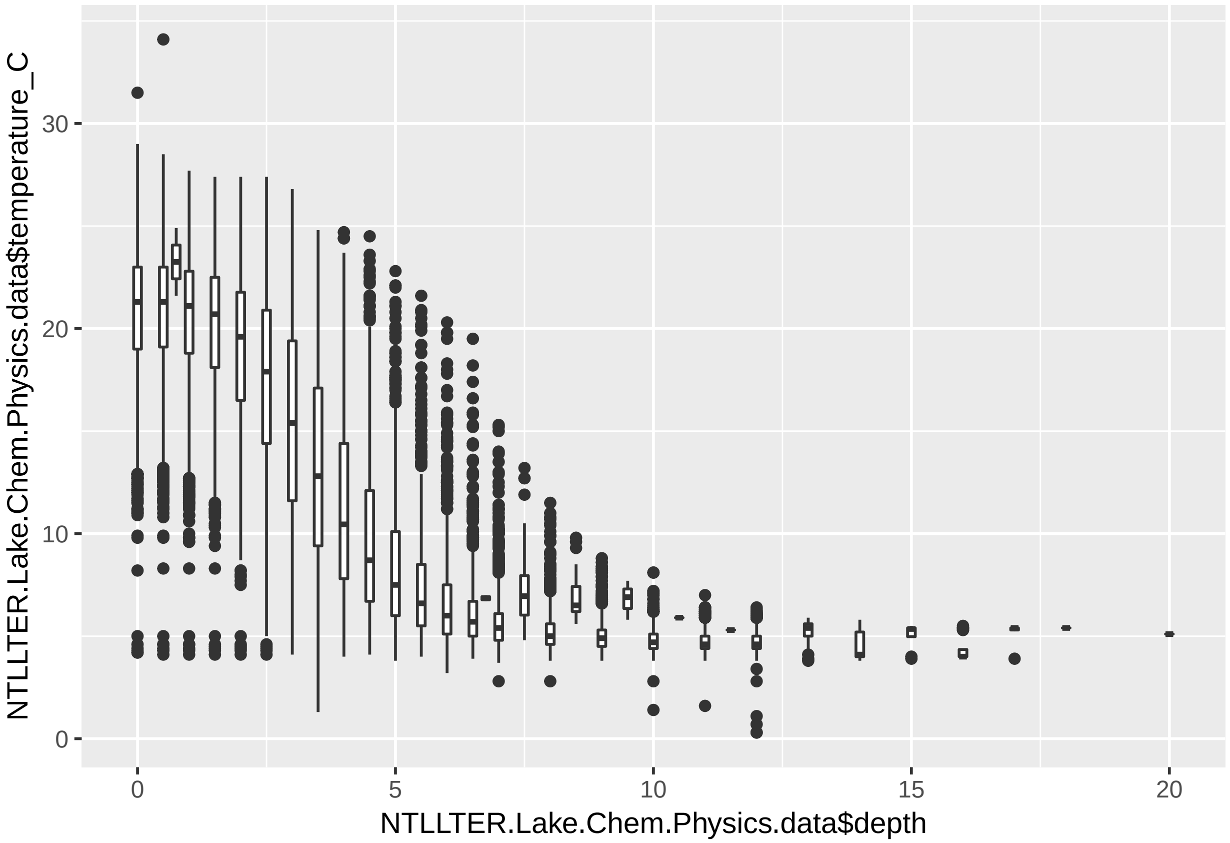
```
# 5 #boxplot of temperature for each lake
ggplot(NTLLTER.Lake.Chem.Physics.data) +
  geom_boxplot(aes(x = NTLLTER.Lake.Chem.Physics.data$lakename,
  y = NTLLTER.Lake.Chem.Physics.data$temperature_C))
```

```
## Warning: Removed 3858 rows containing non-finite values (stat_boxplot).
```



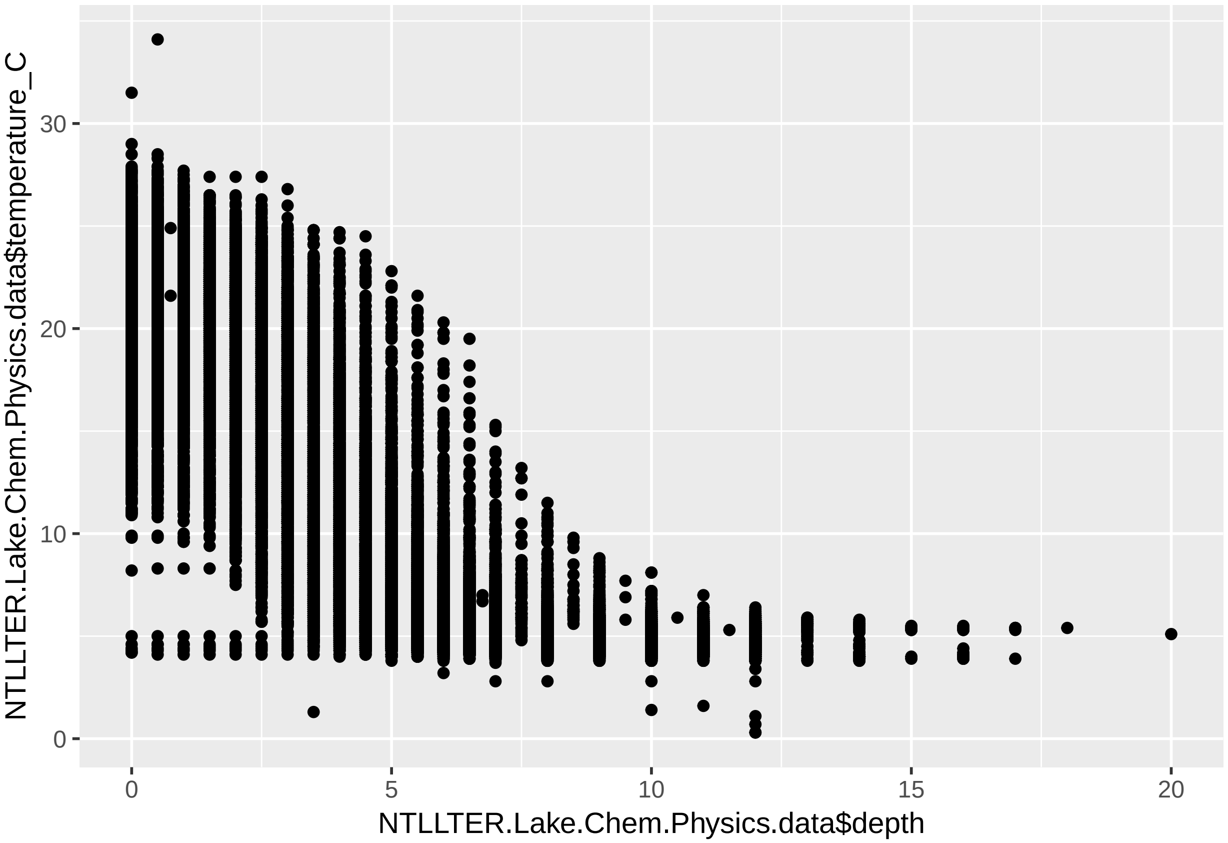
```
# 6 #boxplot of temperature based on depth, with depth divided into 0.25m increments
ggplot(NTLLTER.Lake.Chem.Physics.data) +
  geom_boxplot(aes(x = NTLLTER.Lake.Chem.Physics.data$depth,
  y = NTLLTER.Lake.Chem.Physics.data$temperature_C,
  group = cut_width(NTLLTER.Lake.Chem.Physics.data$depth, 0.25)))
```

```
## Warning: Removed 3858 rows containing non-finite values (stat_boxplot).
```



```r
# 7 #scatterplot of temperature by depth
ggplot(NTLLTER.Lake.Chem.Physics.data) +
  geom_point(aes(x = NTLLTER.Lake.Chem.Physics.data$depth,
  y = NTLLTER.Lake.Chem.Physics.data$temperature_C))
```

```
## Warning: Removed 3858 rows containing missing values (geom_point).
```



# 5) Form questions for further data analysis

What did you find out about your data from the basic summaries and graphs you made? Describe in 4-6 sentences.

> ANSWER: The temperature of the lakes varies, with mean temperatures between 7 degrees Celcius and 17 degrees Celcius. The greatest count of temperatures was between 0 degrees and 5 degrees Celcius. Peter and Paul lakes had high numbers of between about 5 - 7 degrees Celcius and the frequency polygon shows this spike. Generally, as the depth of the lake increases, the temperature is very low. Lake depths were looked at in varying increments, with depths under 1 m broken up into segments of 0.25 m, and lakes with depths above 1 m were generally intergers.

What are 3 further questions you might ask as you move forward with analysis of this dataset?

> ANSWER 1: Does the relationship between temperature and depth change if you look at various sub-groups (i.e. amount of DO, or iradiance water)?

> ANSWER 2: What other variables might influence the relationship between temperature and depth?

> ANSWER 3: Is there a relationship between depth and amount of dissolved oxygen?