Assignment 3: Data Exploration

Rachel Bash

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.

```r
#getwd()
suppressMessages(library(tidyverse))
NTL.Lakes <- read.csv("../Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv")
#thinks its in the Assignments folder, so two .. will tell you go to back out to
#main folder and then into the Data folder.
#View(NTL.Lakes)
```

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 Carbon data and the Nutrients data had different sampling years. This is a good thing to keep in mind if we have to deal with dates. There are also different sampling methods depending on the date when the samples were taken (specifically the Nutrients section). The different sampling methods could affect the data in some way. Lastly, I noticed that there were a lot of variables–the various samples with which frequency the data was collected.

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
dim(NTL.Lakes)
```

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

```
# 2
class(NTL.Lakes)
```

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

```
# 3
head(NTL.Lakes,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
str(NTL.Lakes)
```

```
## 'data.frame':    38614 obs. of  11 variables:
##  $ lakeid         : Factor w/ 9 levels "C","E","H","L",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ lakename       : Factor w/ 9 levels "Central Long Lake",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ year4          : int  1984 1984 1984 1984 1984 1984 1984 1984 1984 1984 ...
##  $ daynum         : int  148 148 148 148 148 148 148 148 148 148 ...
##  $ sampledate     : Factor w/ 1712 levels "10/1/07","10/1/93",..: 134 134 134 134 134 134 134 134 13
##  $ depth          : num  0 0.25 0.5 0.75 1 1.5 2 3 4 5 ...
##  $ temperature_C  : num  14.5 NA NA NA 14.5 NA 14.2 11 7 6.1 ...
##  $ dissolvedOxygen: num  9.5 NA NA NA 8.8 NA 8.6 11.5 11.9 2.5 ...
##  $ irradianceWater: num  1750 1550 1150 975 870 610 420 220 100 34 ...
##  $ irradianceDeck : num  1620 1620 1620 1620 1620 1620 1620 1620 1620 1620 ...
##  $ comments       : Factor w/ 2 levels "DO Probe bad - Doesn't go to zero",..: NA NA NA NA NA NA NA
```

```
#lakename = factor; sampledate = factor; depth = number; temperature = number
# 5
summary(NTL.Lakes$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
```

```
summary(NTL.Lakes$depth)
```

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

```
summary(NTL.Lakes$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.

```
NTL.Lakes$sampledate <- as.Date(NTL.Lakes$sampledate, format = "%m/%d/%y")
class(NTL.Lakes$sampledate)
```

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

```
head(NTL.Lakes$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?

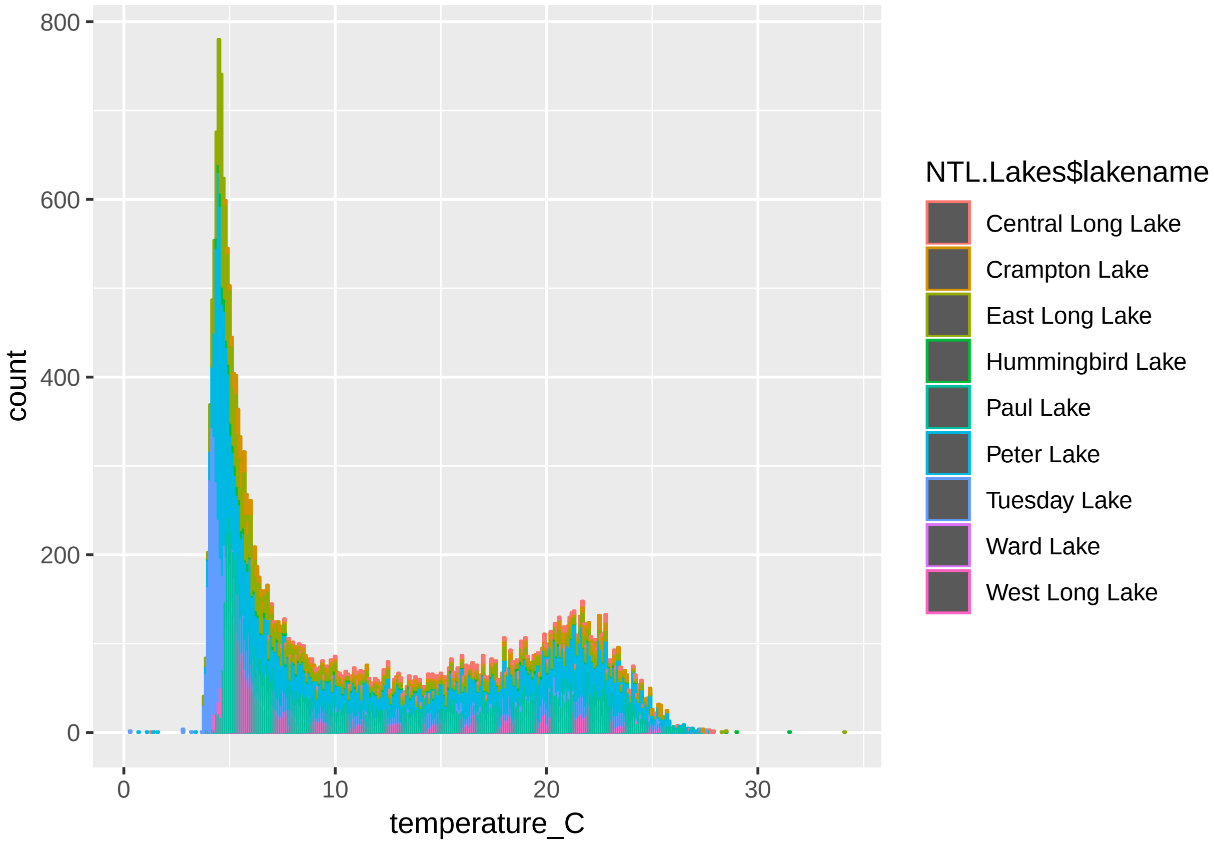
> ANSWER: No. The NAs are dispersed throughout the dataset and for each row, there are some data but not others. If there is good data in some rows, it would be a waste to remove the NAs because you would be removing the entire row of data. Because they are scattered throughout the dataset, it is okay to leave them in. This is different from the example we had in class where entire columns were NA up to a certain date.

4) Explore your data graphically

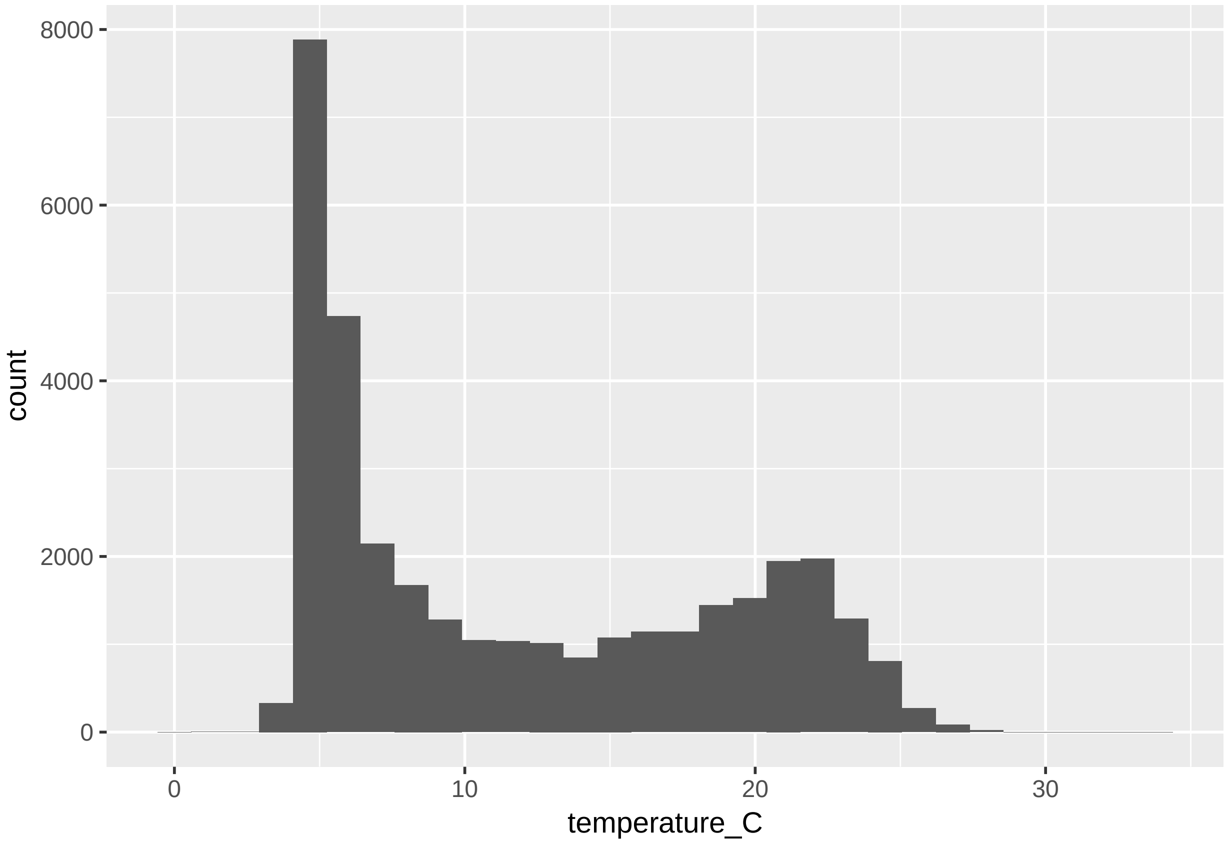
Write R commands to display graphs depicting:

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

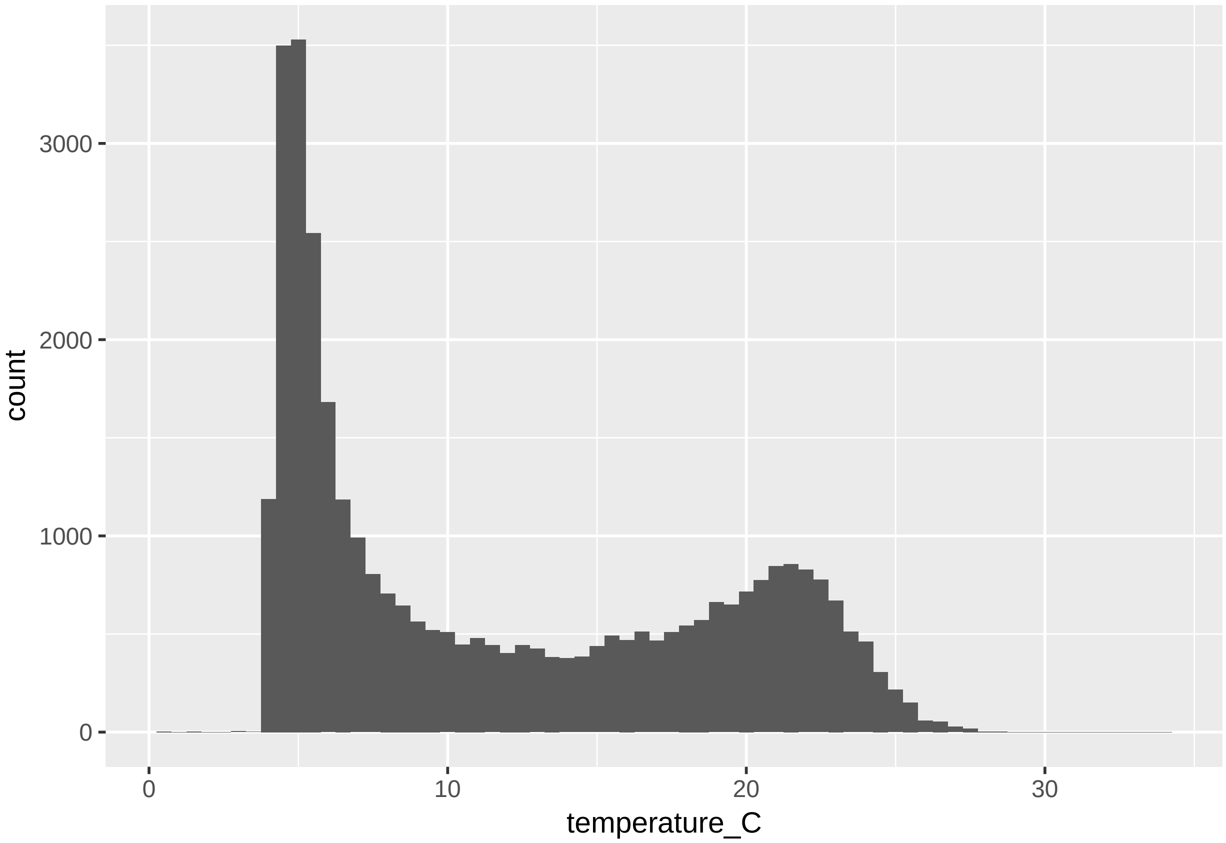
```
# 1
ggplot(NTL.Lakes) +
  geom_bar(aes(x=temperature_C, color= NTL.Lakes$lakename))
```



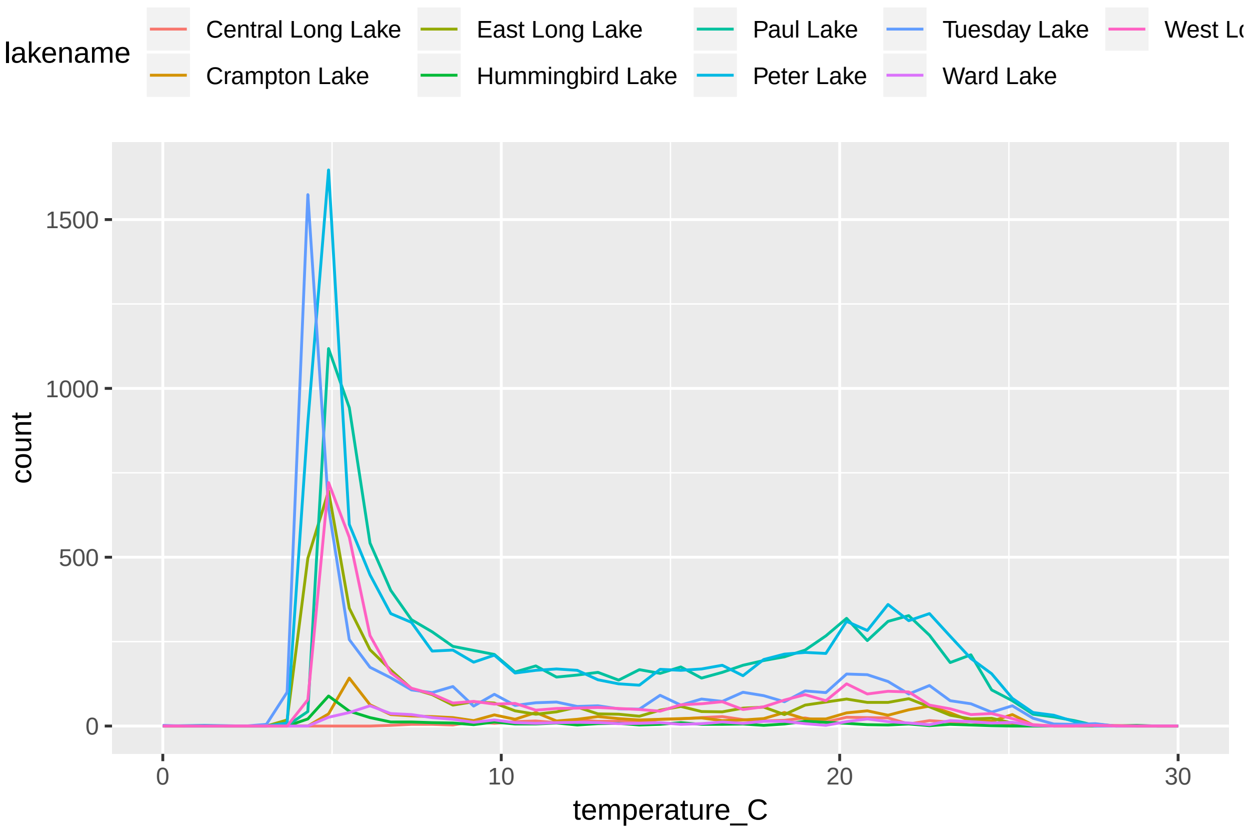
```
# 2
ggplot(NTL.Lakes) +
  geom_histogram(aes(x = temperature_C))
```



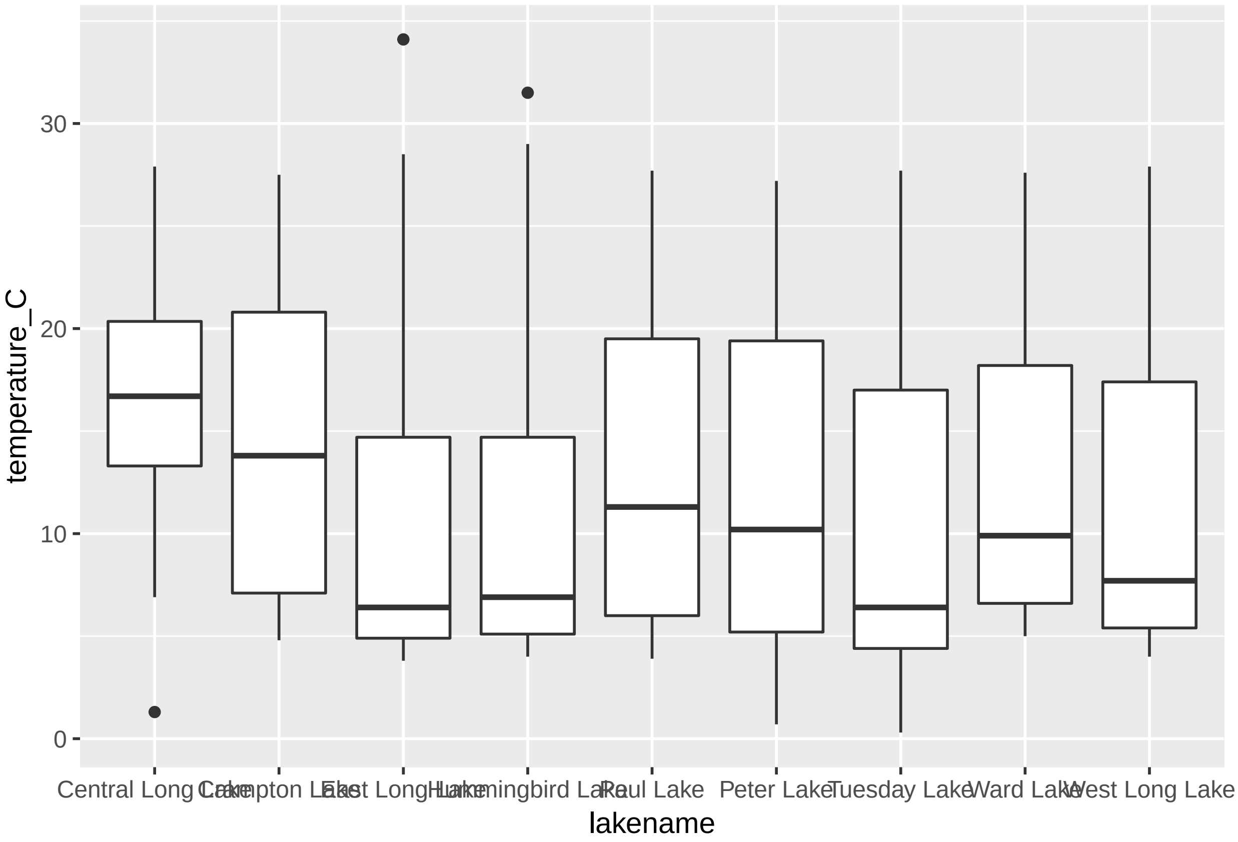
```
# 3
ggplot(NTL.Lakes) +
  geom_histogram(aes(x = temperature_C), binwidth=.5)
```



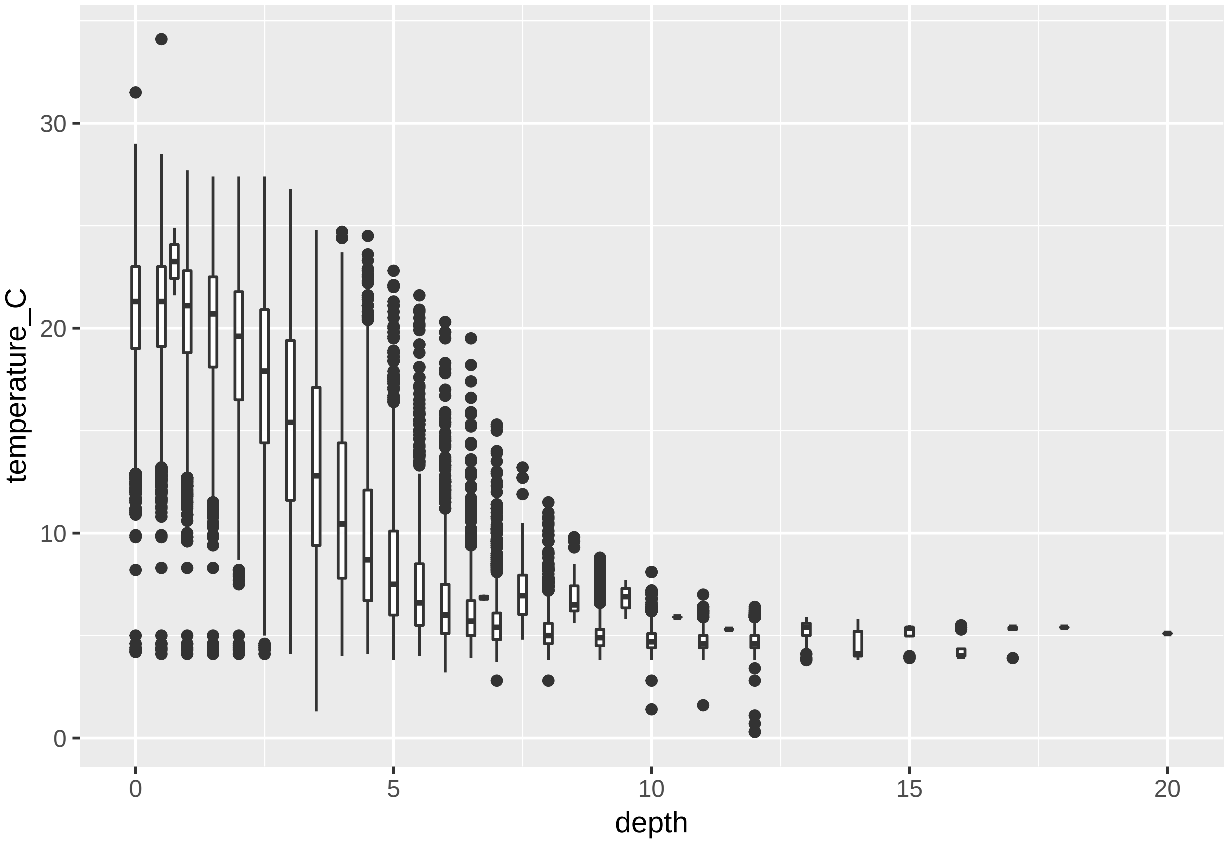
```
# 4
ggplot(NTL.Lakes) +
  geom_freqpoly(aes(x = temperature_C, color = lakename),
  bins = 50) +
  scale_x_continuous(limits = c(0, 30)) +
  theme(legend.position = "top")
```



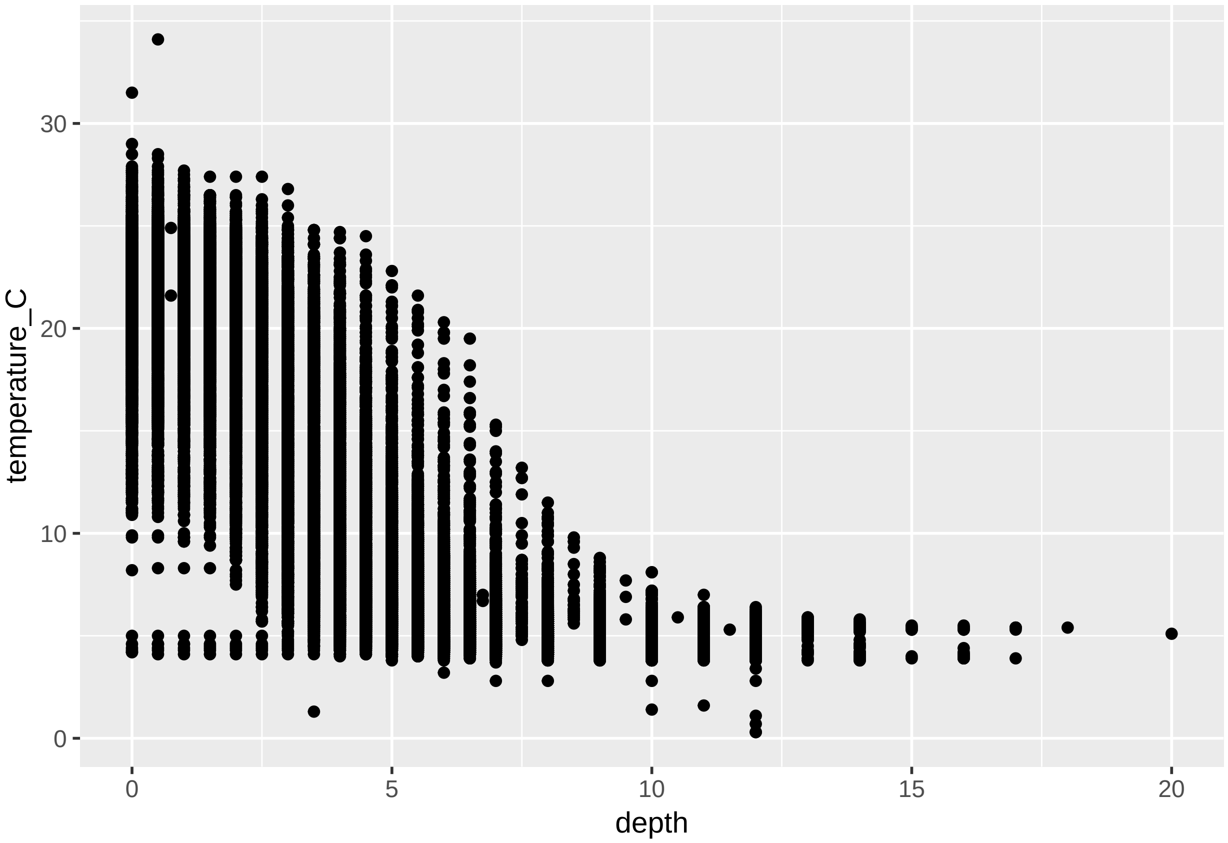
```
# 5
ggplot(NTL.Lakes) +
  geom_boxplot(aes(x = lakename, y = temperature_C))
```



```
# 6
ggplot(NTL.Lakes) +
  geom_boxplot(aes(x= depth, y = temperature_C,
                   group = cut_width(depth,0.25)))
```



```
# 7
ggplot(NTL.Lakes) +
  geom_point(aes(x = depth, y = temperature_C))
```



## 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: Well, first I know that some ways of depicting data summaries are better than others. For example, the first bar graph isn't the best way to display temperature data for each lake because the bins are so small. The frequency polygon graph does a better job of showing that each lake has different variances of temperature. I also learned that, at a shallower depth, there is a bigger range of temperature differences, and the median temperature is higher at a shallower depth. The majority of temperatures that were measured for all lakes were around 5 or 6 degrees Celsius, but there histogram shows that there is a bimodel nature to the curve of the temperature distributions. There is a peak at 5-6 degrees Celsius and also a smaller peak at around 22-23 degrees Celsius.

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

> ANSWER 1: How does dissolved Oxygen differ with temperature?

> ANSWER 2: Does dissolved Oxygen vary across different lakes? What is the range of values that each lake has?

> ANSWER 3: Are depth and temperature significantly correlated? To what degree?