Assignment 3: Data Exploration

Walker Grimshaw

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.

```
## check working directory
getwd() ## in Assignments folder
```

```
## [1] "/Users/walkergrimshaw/Documents/Duke/Courses/Spring_2019/Environmental_Data_Analytics/Assignmen
```

```
## load tidyverse package, try suppressing messages to allow knitting
suppressMessages(library(tidyverse))
## read in Lakes data from Data folder
## If working directory defaults to assignments folder, check the knitting directory
Lake_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: 1) There are three related files with North Temperate Lakes data containing Carbon data, Physical and Chemical Limnology Data, and Nutrient Data. 2) The Carbon and Physical and Chemical Limnology Data collected span the time interval from 1984 to 2016, and the Nutrient data span the interval from 1991 to 2016. 3) The carbon data were collected at 14 separate sites, while the nutrient and physical and chemical were collected from roughly the deepest point of each lake.

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(Lake_Data)
```

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

```
# 2
class(Lake_Data)
```

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

```
# 3
head(Lake_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(Lake_Data$lakename)
```

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

```
class(Lake_Data$sampledate)
```

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

```
class(Lake_Data$depth)
```

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

```
class(Lake_Data$temperature_C)
```

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

```
# 5
summary(Lake_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
```

```
summary(Lake_Data$depth)
```

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

```
summary(Lake_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.

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

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

```
## first 10 rows of date column
head(Lake_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: There are many days when irradiance was not measured, but temperature and DO were still measured. There is still useful data analysis to be performed on these data points, but all that information would be lost if we removed all rows with any NAs present. For that reason, I do not think we should remove the NAs.

4) Explore your data graphically

Write R commands to display graphs depicting:

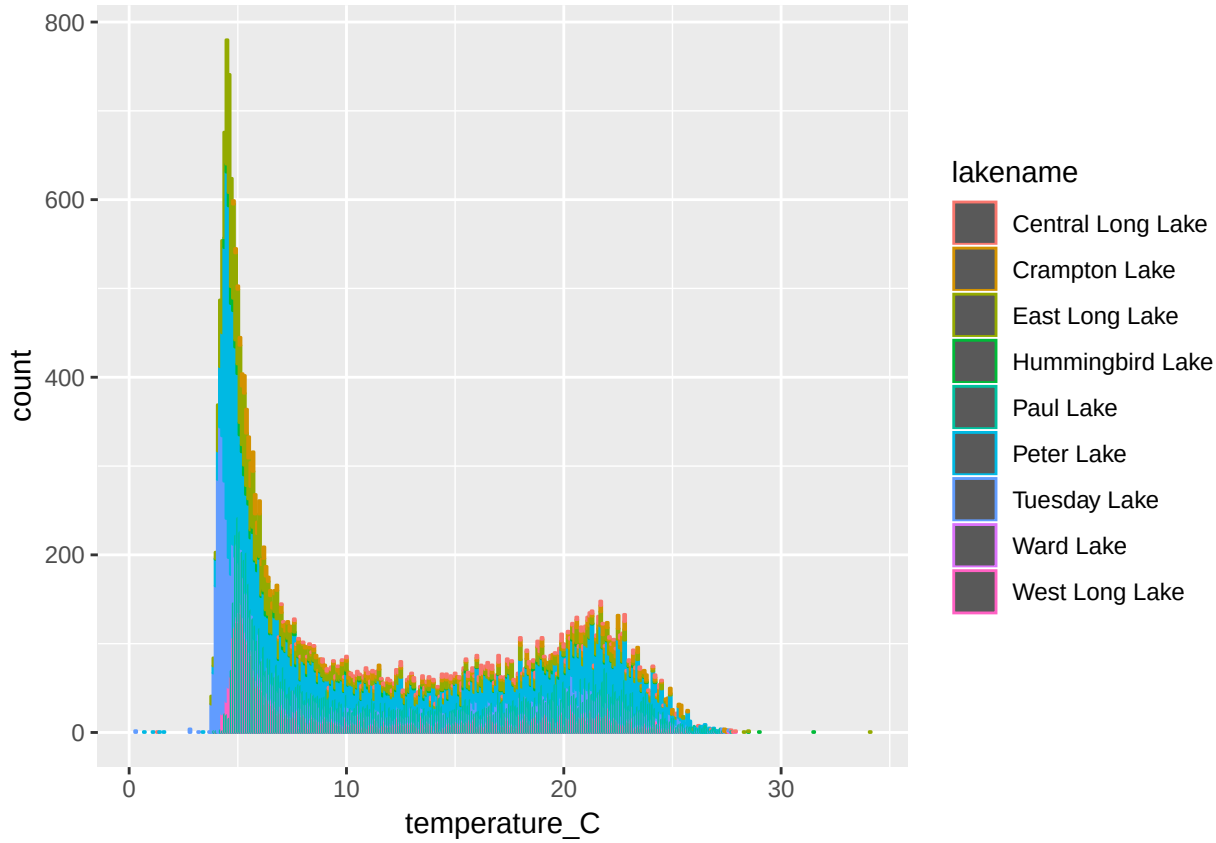
1. Bar chart of temperature counts for each lake
2. Histogram of count distributions of temperature (all temp measurements together)
3. Change histogram from 2 to have a different number or width of bins
4. Frequency polygon of temperature for each lake. Choose different colors for each lake.
5. Boxplot 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 Bar chart of temperature counts

ggplot(Lake_Data) +
  geom_bar(aes(x = temperature_C, color = lakename))
```

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

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

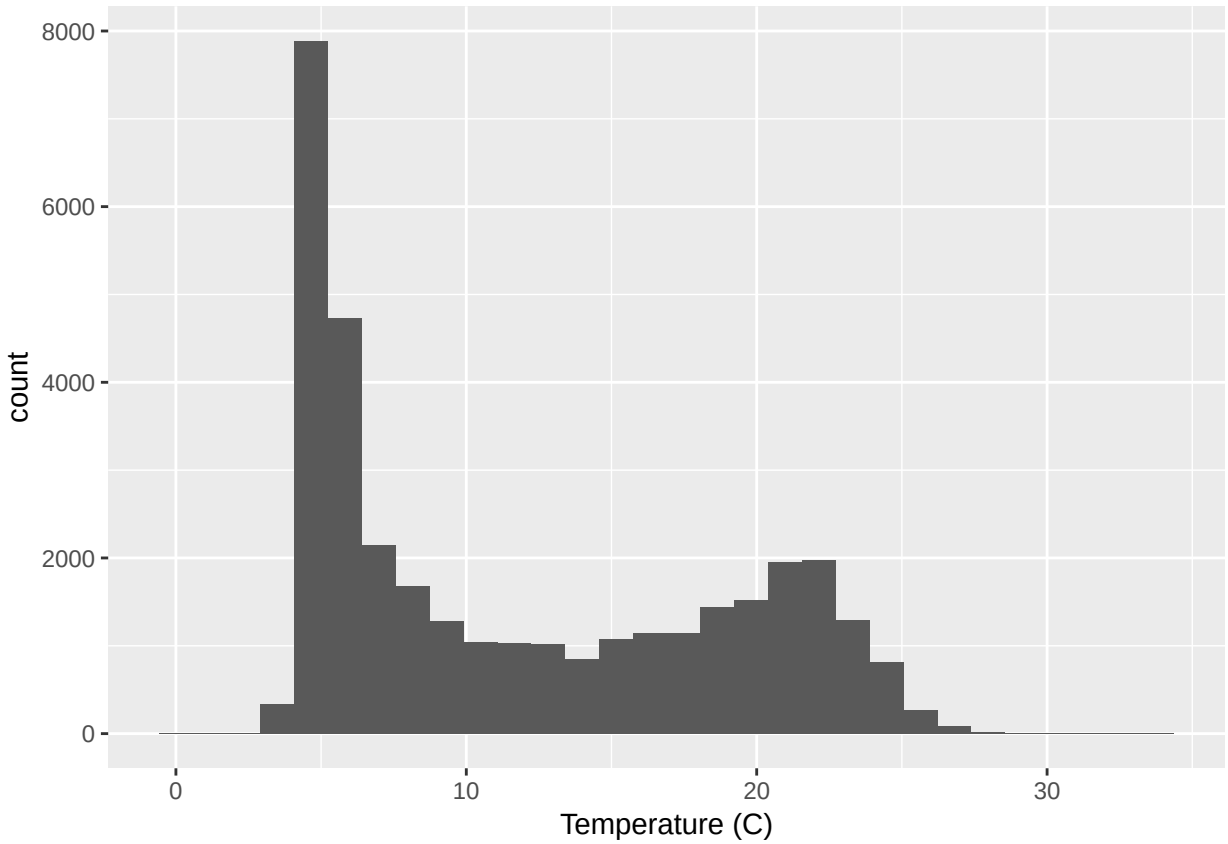



```
# 2 Histogram of temperature across all lakes

ggplot(Lake_Data) +
  geom_histogram(aes(x = temperature_C)) +
  xlab("Temperature (C)")
```

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

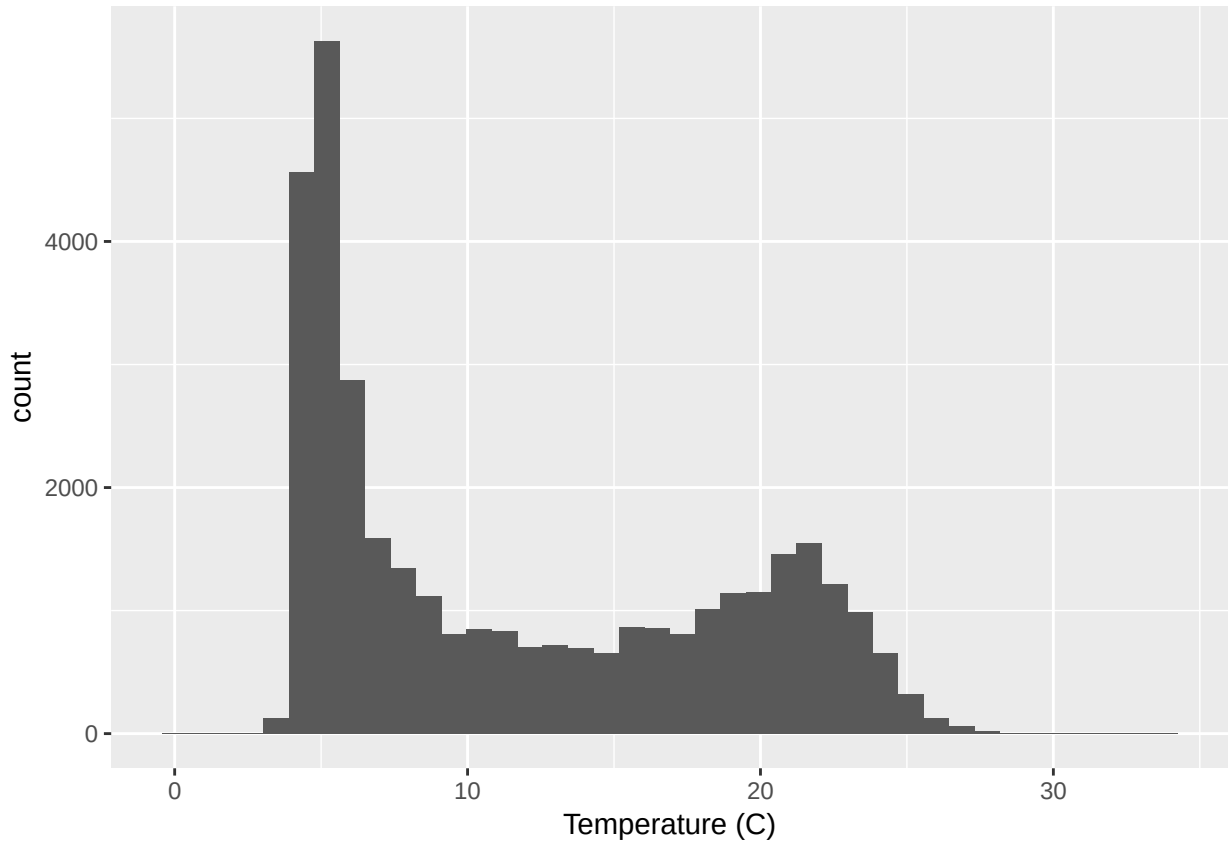
```
## Warning: Removed 3858 rows containing non-finite values (stat_bin).
```




```
# 3 Histogram, same as #2, but with 40 bins

ggplot(Lake_Data) +
  geom_histogram(aes(x = temperature_C), bins = 40) +
  xlab("Temperature (C)")
```

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

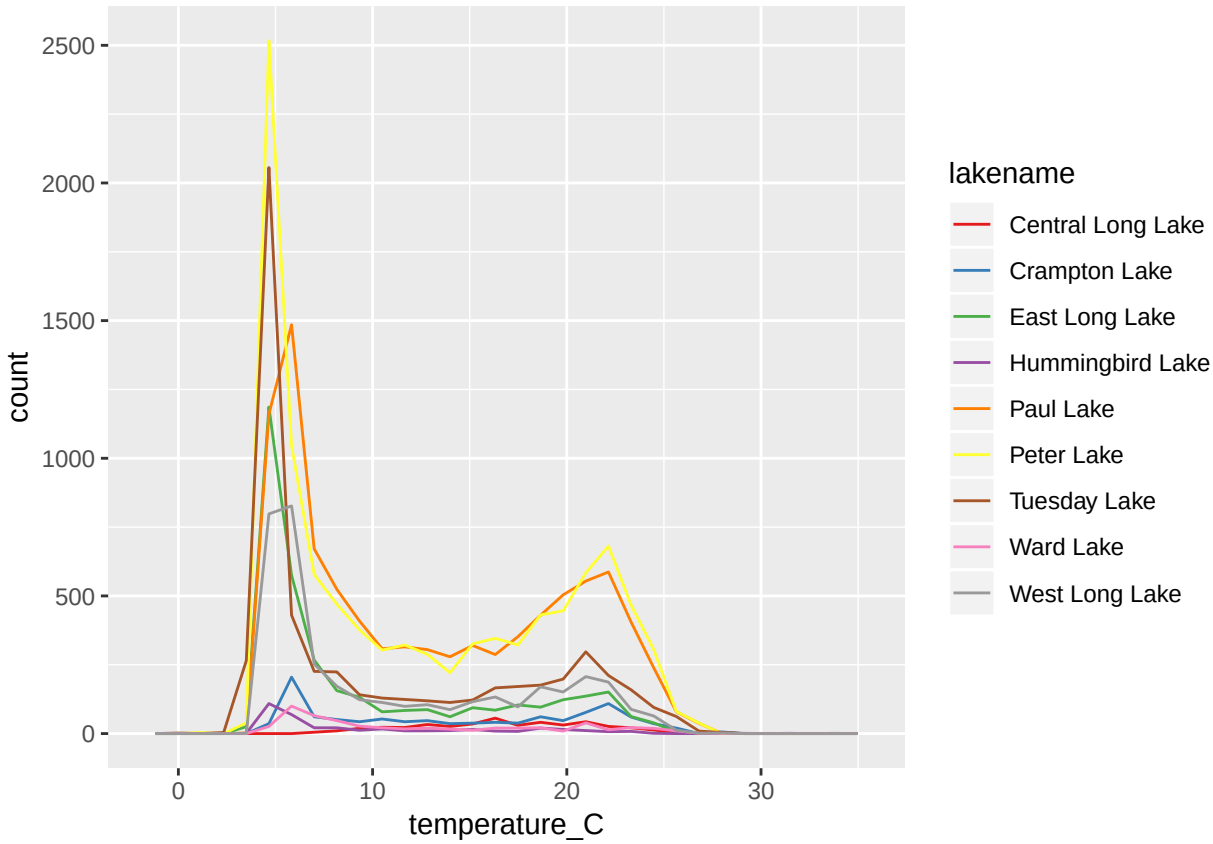



```
# 4 Frequency polygon of temp for each lake

ggplot(Lake_Data) +
  geom_freqpoly(aes(x = temperature_C, color = lakename)) +
  scale_color_brewer(palette="Set1")
```

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

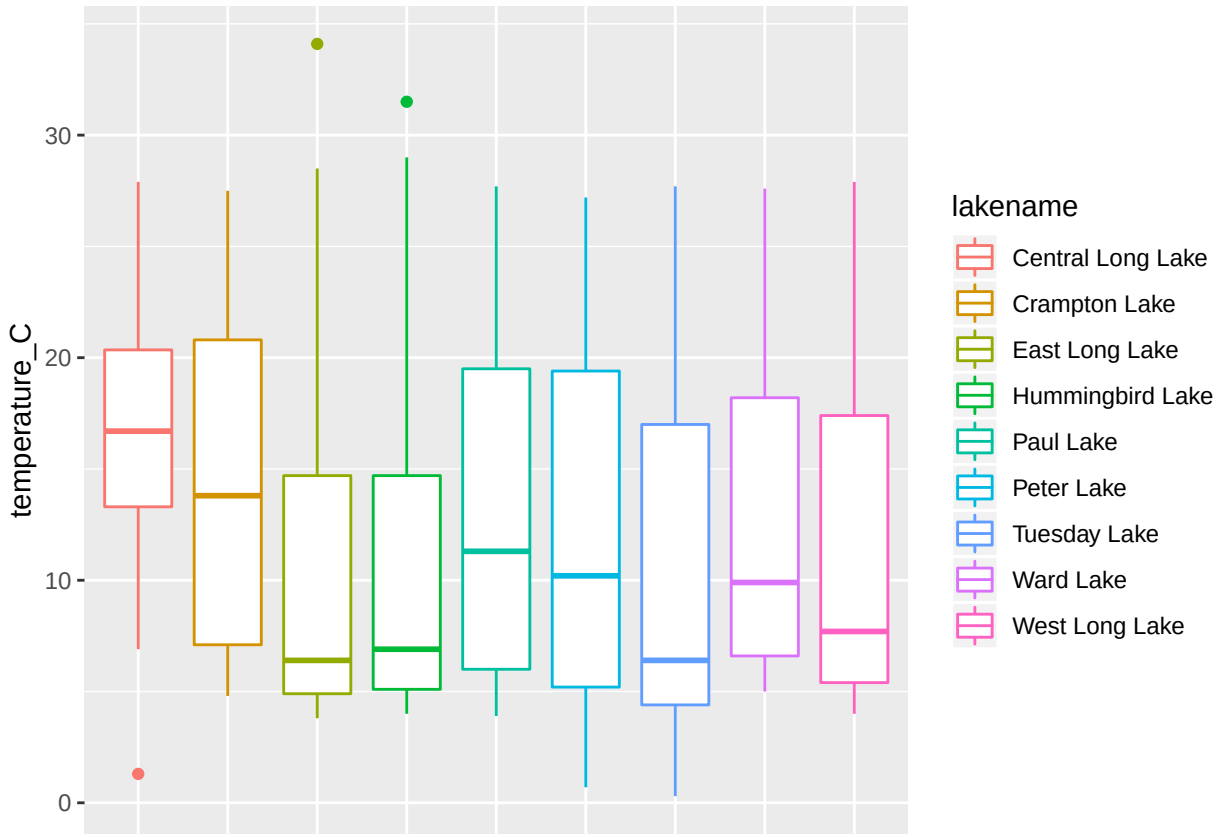
```
## Warning: Removed 3858 rows containing non-finite values (stat_bin).
```




```
# 5 Boxplot of temp for each lake

ggplot(Lake_Data) +
  geom_boxplot(aes(x = lakename, y = temperature_C, color = lakename)) +
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank(),
        axis.title.x = element_blank())
```

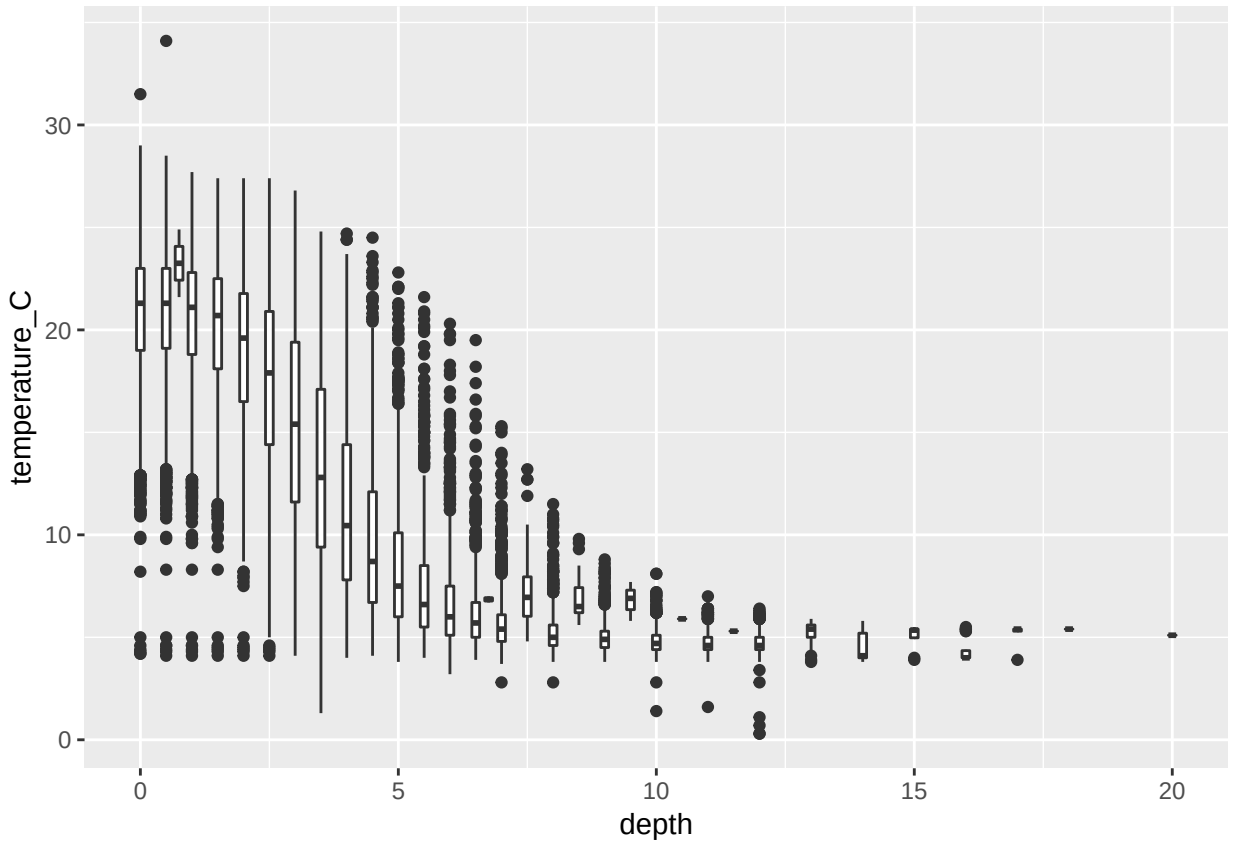
```
## Warning: Removed 3858 rows containing non-finite values (stat_boxplot).
```




```
# 6

ggplot(Lake_Data) +
  geom_boxplot(aes(x = depth, y = temperature_C, group = cut_width(depth, 0.25)))
```

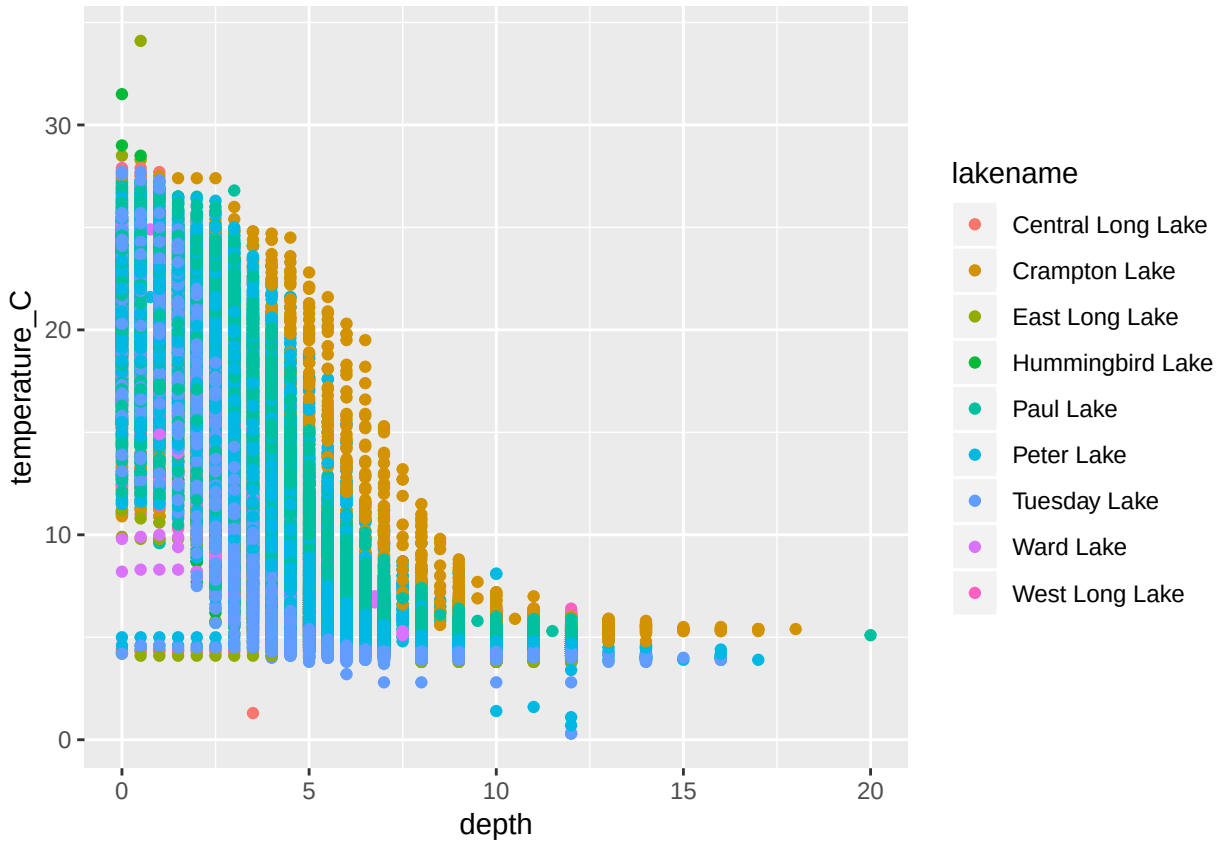
```
## Warning: Removed 3858 rows containing non-finite values (stat_boxplot).
```




```
# 7 scatterplot depth vs temp

ggplot(Lake_Data) +
  geom_point(aes(x = depth, y = temperature_C, color = lakename))
```

```
## 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: All lakes appear from the frequency polygons to follow roughly the same temperature distribution, but the boxplots show that there is actually a fair amount of variation in temperature distributions between lakes. Lakes such as Paul Lake and Peter Lake have much more temperature data than Ward Lake and Hummingbird Lake, but we do not know if this is due to more monitoring or because of lake size. Water temperature decreases with water depth, though the temperature of the water levels off around 5 C even when very deep. The temperature data collected have a bimodal distribution, with a peak near 5 C and another peak near 22 C.

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

> ANSWER 1: What is the relationship between irradiance and water temperature, and does this vary by lake?
>
> ANSWER 2: What is the relationship between irradiance and depth, and does this vary by lake?
>
> ANSWER 3: How do any of the variables, especially dissolved oxygen, change over time? What are the impacts of this change on the ecosystem and what might cause any changes, if present?