Assignment 3: Data Exploration

Lindsay Roth

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.

```
setwd("~/Documents/MEM 1st Year/Spring 2019/Env_Data_Analytics/Env_Data_Analytics")
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

```
## -- Conflicts ---------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
lake.chem <- read_csv("Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv")
```

```
## Parsed with column specification:
## cols(
##   lakeid = col_character(),
##   lakename = col_character(),
##   year4 = col_double(),
##   daynum = col_double(),
##   sampledate = col_character(),
##   depth = col_double(),
```

```
##   temperature_C = col_double(),
##   dissolvedOxygen = col_double(),
##   irradianceWater = col_double(),
##   irradianceDeck = col_double(),
##   comments = col_logical()
## )
```

```
## Warning: 368 parsing failures.
##   row      col           expected                              actual
## 36649 comments 1/0/T/F/TRUE/FALSE DO Probe bad - Doesn't go to zero 'Data/Raw/NTL-LTER_Lake_Chemistry
## 36651 comments 1/0/T/F/TRUE/FALSE DO Probe bad - Doesn't go to zero 'Data/Raw/NTL-LTER_Lake_Chemistry
## 36653 comments 1/0/T/F/TRUE/FALSE DO Probe bad - Doesn't go to zero 'Data/Raw/NTL-LTER_Lake_Chemistry
## 36654 comments 1/0/T/F/TRUE/FALSE DO Probe bad - Doesn't go to zero 'Data/Raw/NTL-LTER_Lake_Chemistry
## 36655 comments 1/0/T/F/TRUE/FALSE DO Probe bad - Doesn't go to zero 'Data/Raw/NTL-LTER_Lake_Chemistry
## ..... ........ .................. ................................. ..............................
## See problems(...) for more details.
```

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: When the data was accessed, how the data was collected, and the content of the dataset including what chemical parameters are part of the dataset.

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.chem)
```

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

```
# 2
class(lake.chem)
```

```
## [1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
```

```
# 3
head(lake.chem,8)
```

```
## # A tibble: 8 x 11
##   lakeid lakename year4 daynum sampledate depth temperature_C
##   <chr>  <chr>    <dbl>  <dbl> <chr>      <dbl>         <dbl>
## 1 L      Paul La~  1984    148 5/27/84     0             14.5
## 2 L      Paul La~  1984    148 5/27/84     0.25          NA
## 3 L      Paul La~  1984    148 5/27/84     0.5           NA
## 4 L      Paul La~  1984    148 5/27/84     0.75          NA
## 5 L      Paul La~  1984    148 5/27/84     1             14.5
```

```
## 6 L      Paul La~  1984    148 5/27/84     1.5           NA
## 7 L      Paul La~  1984    148 5/27/84     2             14.2
## 8 L      Paul La~  1984    148 5/27/84     3             11
## # ... with 4 more variables: dissolvedOxygen <dbl>, irradianceWater <dbl>,
## #   irradianceDeck <dbl>, comments <lgl>
```

```
# 4
class(lake.chem$lakename)
```

```
## [1] "character"
```

```
class(lake.chem$sampledate)
```

```
## [1] "character"
```

```
class(lake.chem$depth)
```

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

```
class(lake.chem$temperature_C)
```

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

```
# 5
summary(lake.chem$lakename)
```

```
##    Length     Class      Mode
##     38614 character character
```

```
summary(lake.chem$depth)
```

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

```
summary(lake.chem$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.chem$sampledate <- as.Date(lake.chem$sampledate, format = "%m/%d/%y")
head(lake.chem$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: I would like to remove the NAs from the dataset because they will alter the accuracy of the statistical analysis for continuous variables such as temperature. However, since the column named "comments" is completely made up of NAs, the entire dataset would be removed if I removed the rows with NAs.

4) Explore your data graphically

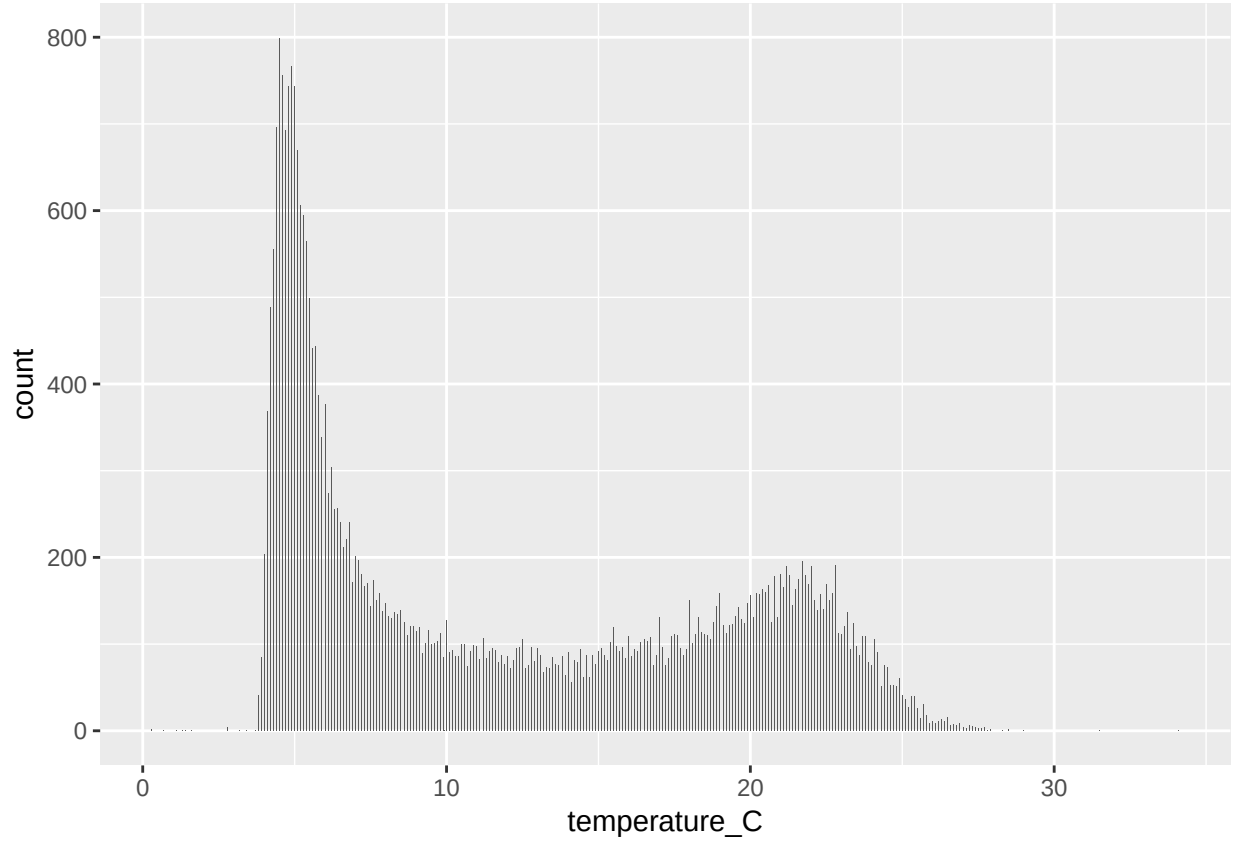
Write R commands to display graphs depicting:

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
ggplot(lake.chem) +
  geom_bar(aes(x = temperature_C))
```

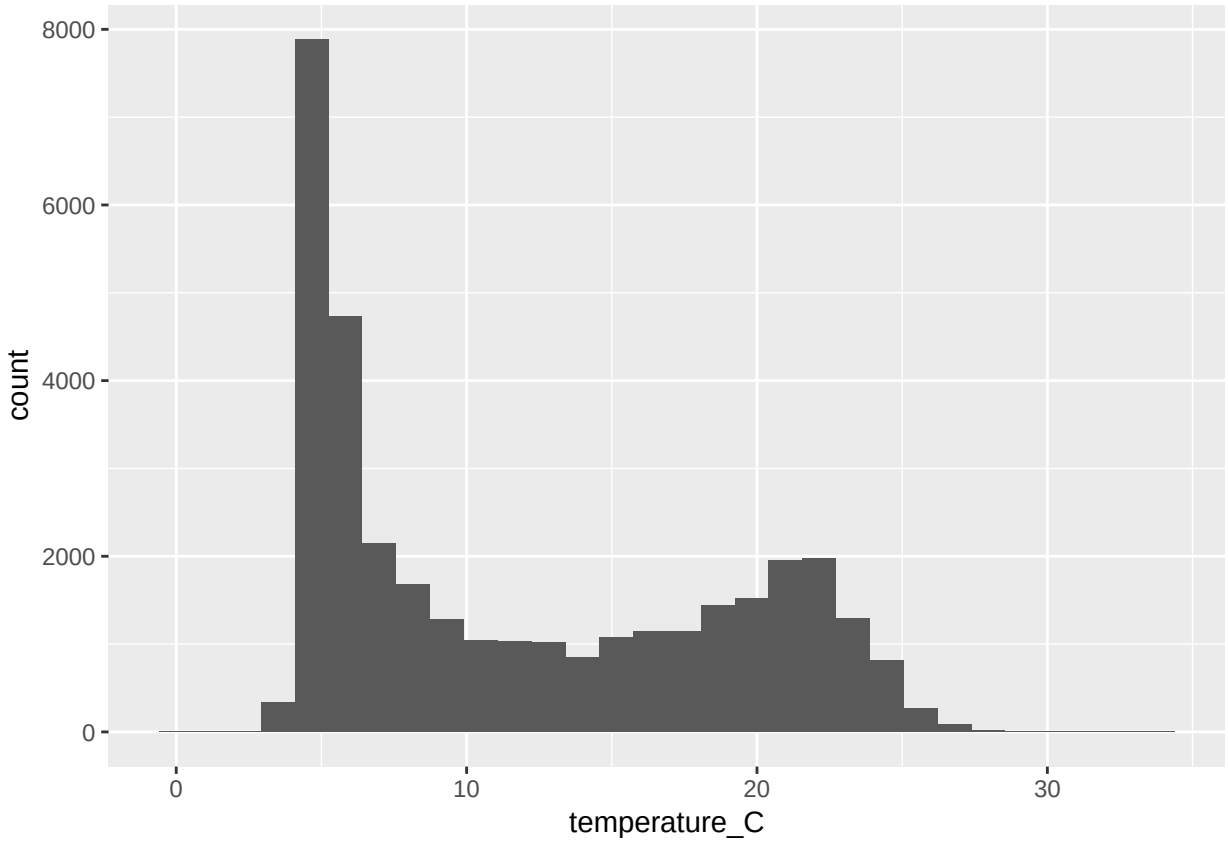
```
## Warning: Removed 3858 rows containing non-finite values (stat_count).
```




```
# 2
ggplot(lake.chem) +
  geom_histogram(aes(x = temperature_C))
```

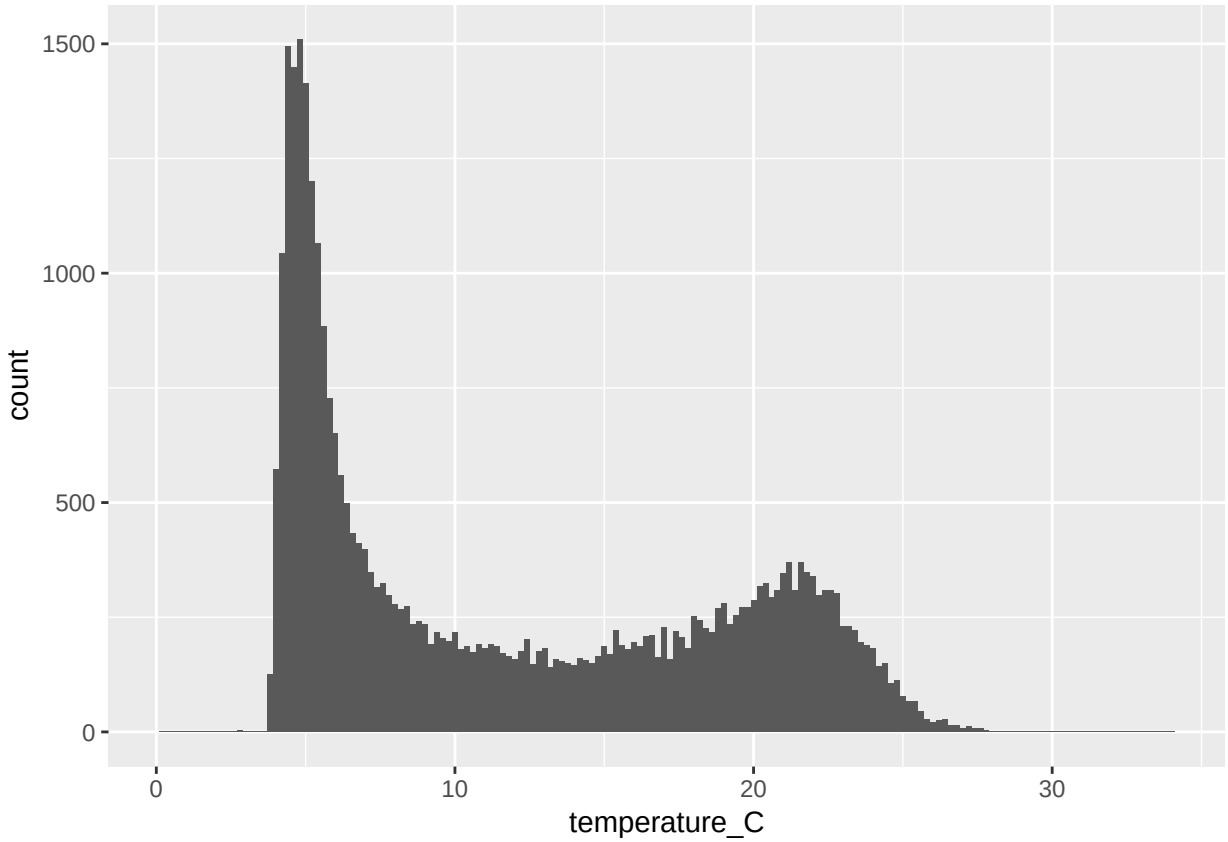
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

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




```
# 3
ggplot(lake.chem) +
  geom_histogram(aes(x = temperature_C), binwidth = 0.2)
```

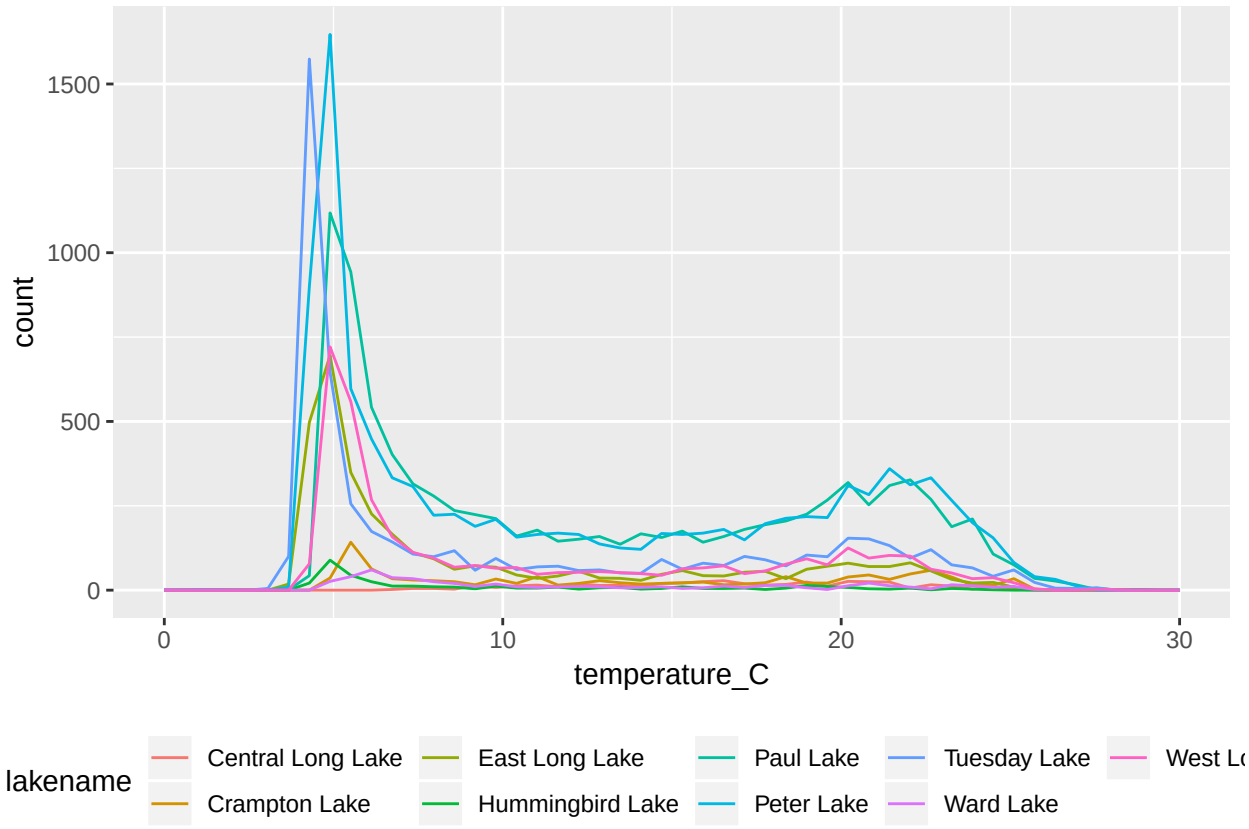
```
## Warning: Removed 3858 rows containing non-finite values (stat_bin).
```




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

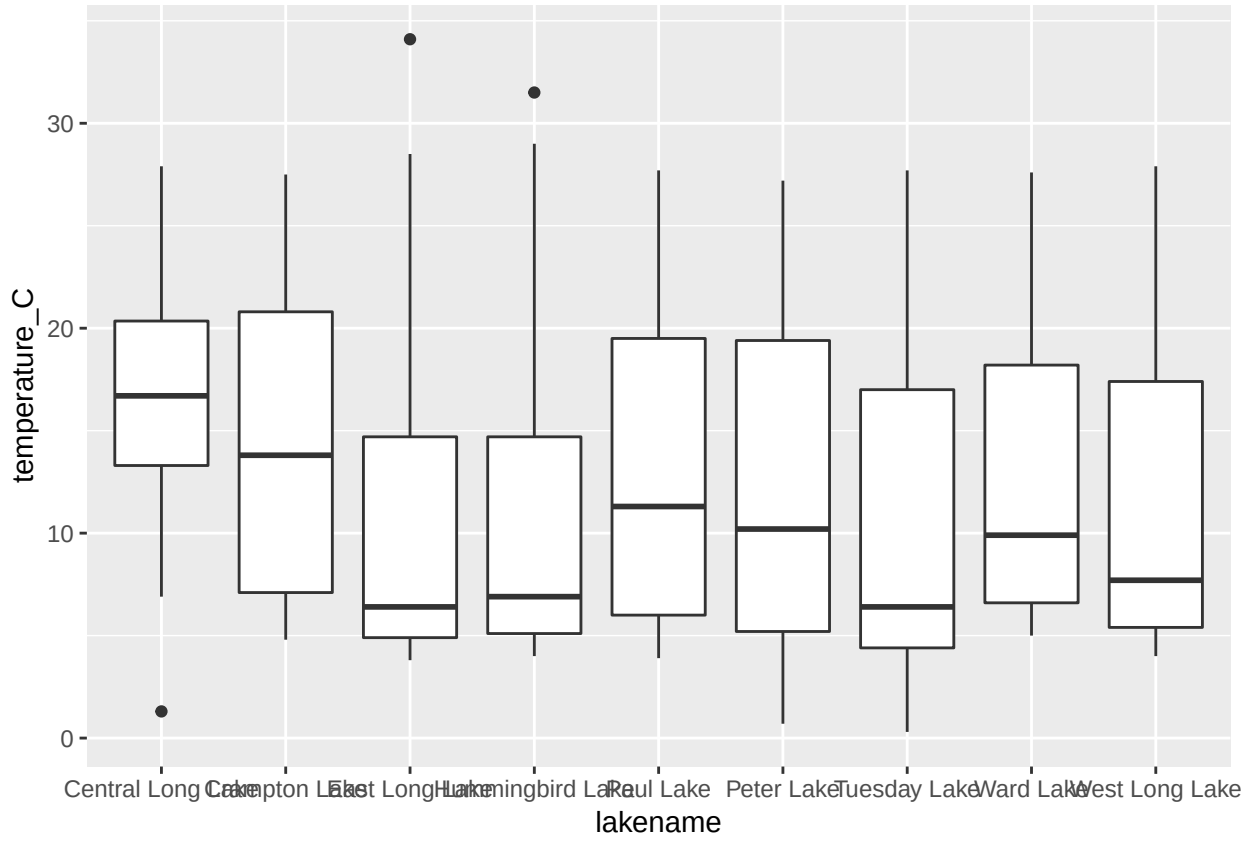
```
## Warning: Removed 3860 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 18 rows containing missing values (geom_path).
```



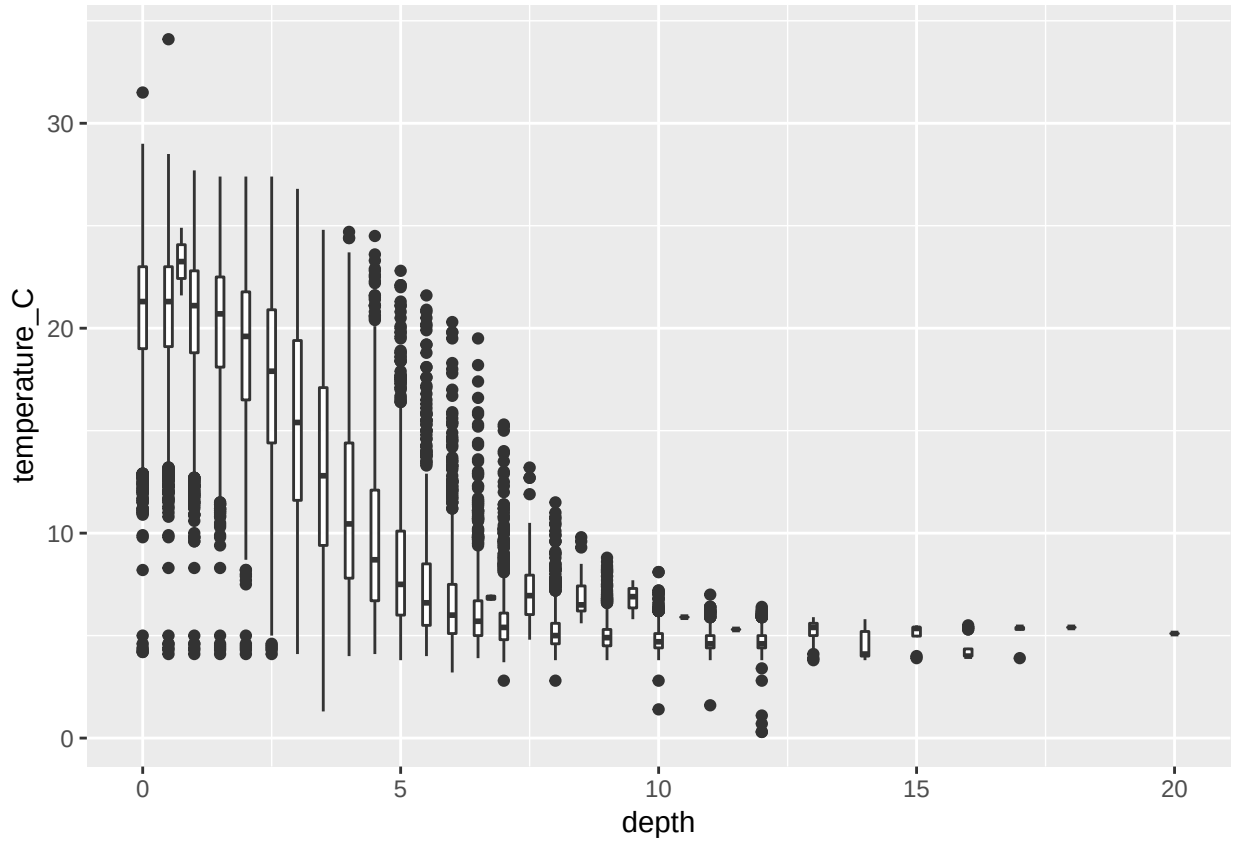

```
# 5
ggplot(lake.chem) +
  geom_boxplot(aes(x = lakename, y = temperature_C))
```

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



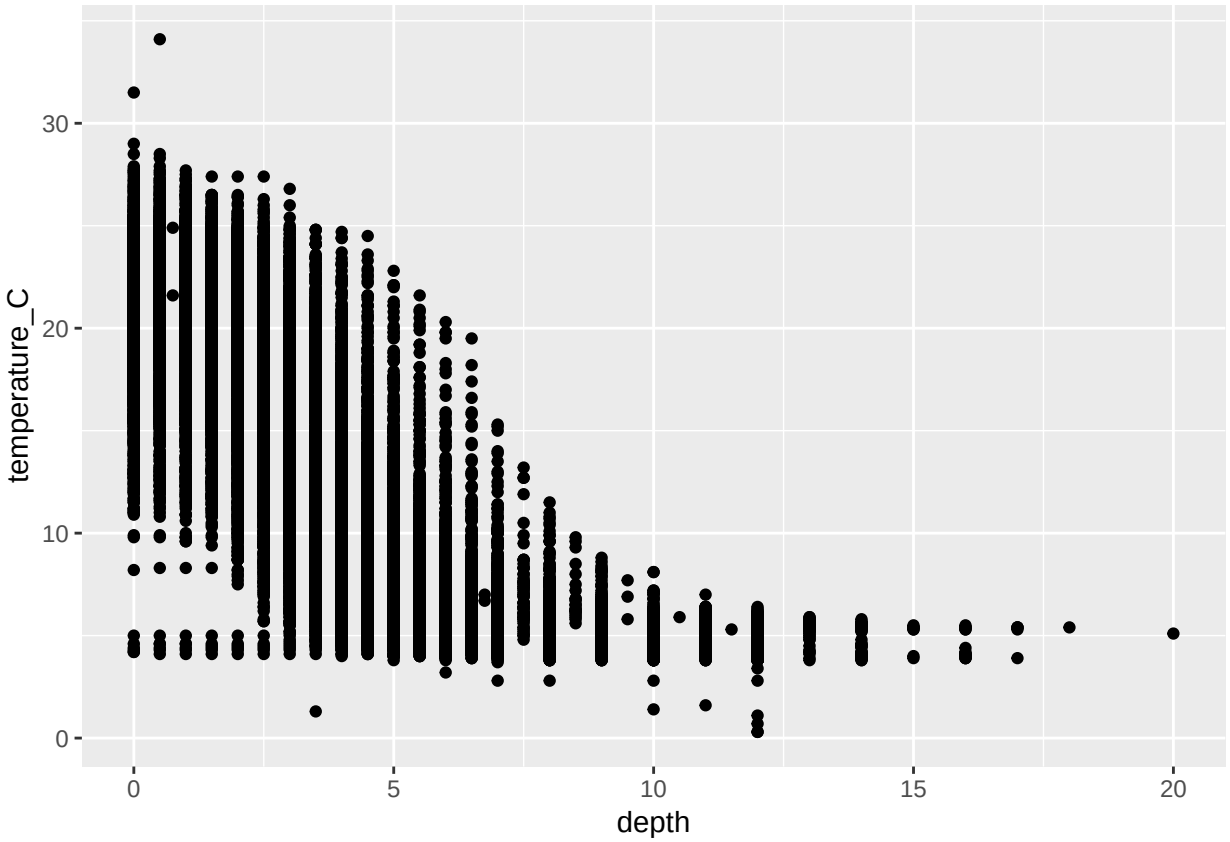

```
# 6
ggplot(lake.chem) +
  geom_boxplot(aes(x = depth, y = temperature_C, group = cut_width(depth, 0.25)))
```

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




```
# 7
ggplot(lake.chem) +
  geom_point(aes(x = depth, y = 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: I learned a lot about the distribution of temperature as a whole, related to depth, and throughout the lakes. First, I noticed that the most frequent temperature measurements were around 5 degrees C, with another large cluster a little higher than 20 degrees C. Next, I noted the differences between the lakes. While Central Long Lake had the highest median temperature, it also had an outlier that was one of the lowest temperature measurements, in contrast to East Long Lake and Hummingbird Lake which had two of the lowest median temperatures but had outliers that were the two highest temperature measurements. Finally, I saw that with increasing depth, temperature variability decreased to being mostly only low temperatures centering around 5 degrees C, whereas more shallower depths had a range of temperatures from about 4 to 30 degrees C

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

> ANSWER 1: Does the location of these lakes influence their average temperatures? (Latitude, elevation)
>
> ANSWER 2: Does proximity urbanization and possible pollutant input impact the temperatures of the lakes?
>
> ANSWER 3: Does the size of the lakes impact the temperatures of the lakes? (volume, surface area)