Assignment 3: Data Exploration

Siying Chen

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 and load tidyverse
getwd()
```

```
## [1] "/Users/Sylvia/Downloads/ENV872/ENV872/Assignments"
```

```
library(tidyverse)
```

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

```
## 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()
```

```
# Import entire dataset and view
NTL_LTER_monitoring <- read.csv("../Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv")
view(NTL_LTER_monitoring)
```

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 are collected from the North Temperate Lakes Long Term Ecological Research website and prepared for this class. Data are obtained from studies on several lakes in the North Temperate Lakes District in Wisconsin, USA. There are two other sets of data, which are carbon data and nutrients data.

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_LTER_monitoring)
```

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

```
# 2
class(NTL_LTER_monitoring)
```

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

```
# 3
head(NTL_LTER_monitoring, 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(NTL_LTER_monitoring$lakename)
```

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

```r
class(NTL_LTER_monitoring$sampledate)
```

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

```r
class(NTL_LTER_monitoring$depth)
```

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

```r
class(NTL_LTER_monitoring$temperature)
```

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

```r
# 5
summary(NTL_LTER_monitoring$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(NTL_LTER_monitoring$depth)
```

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

```r
summary(NTL_LTER_monitoring$temperature)
```

```
##    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
# Change sampledata to date format
NTL_LTER_monitoring$sampledate <- as.Date(NTL_LTER_monitoring$sampledate, format = "%m/%d/%y")

# Check the class of sample date
class(NTL_LTER_monitoring$sampledate)
```

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

```r
# Show the first 10 rows of sampledate column
head(NTL_LTER_monitoring$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: Yes, because if we want to do a statistical analysis on the data, the NAs would affect our result and reduce statistical power.

4) Explore your data graphically

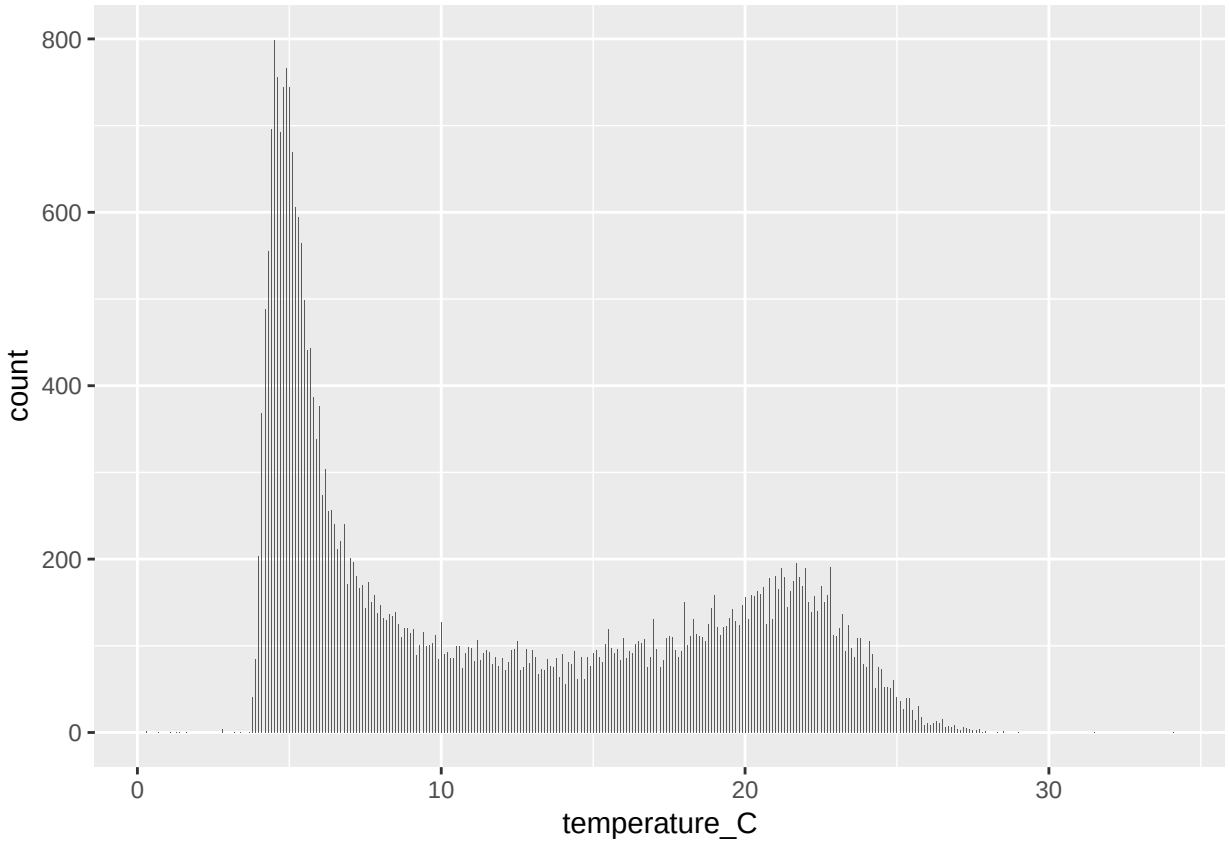
Write R commands to display graphs depicting:

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_LTER_monitoring) +
  geom_bar(aes(x = temperature_C))
```

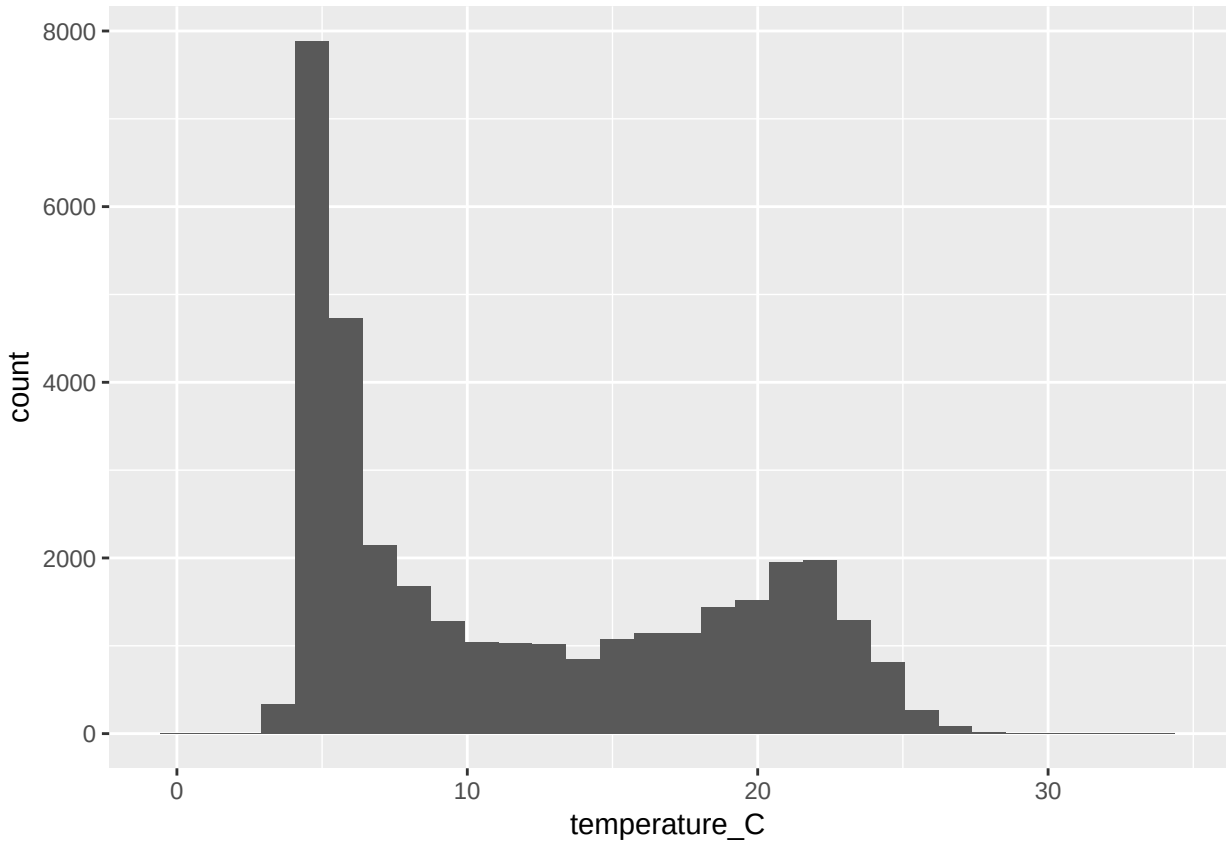
```
## Warning: Removed 3858 rows containing non-finite values (stat_count).
```




```
# 2
ggplot(NTL_LTER_monitoring) +
  geom_histogram(aes(x = temperature_C))
```

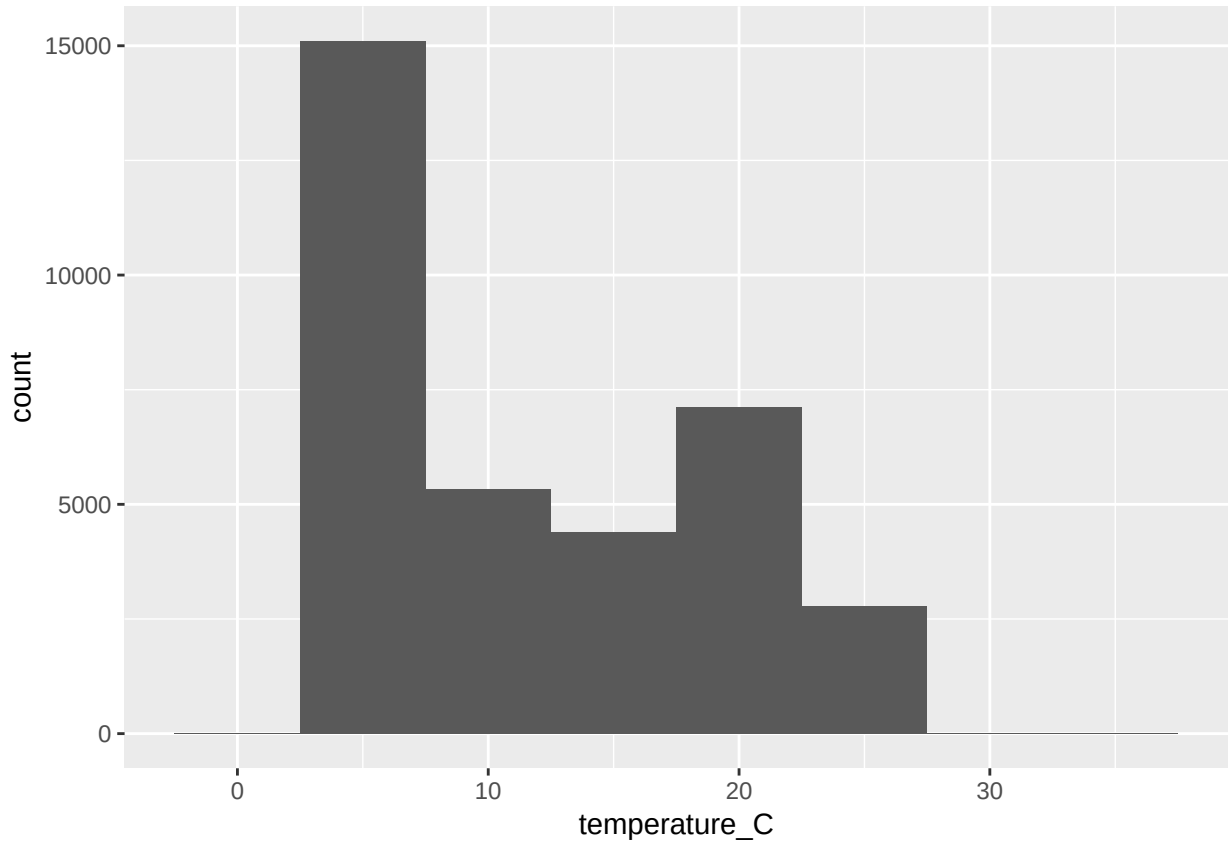
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

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




```
# 3
ggplot(NTL_LTER_monitoring) +
  geom_histogram(aes(x = temperature_C), binwidth = 5) # binwidth of 5
```

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

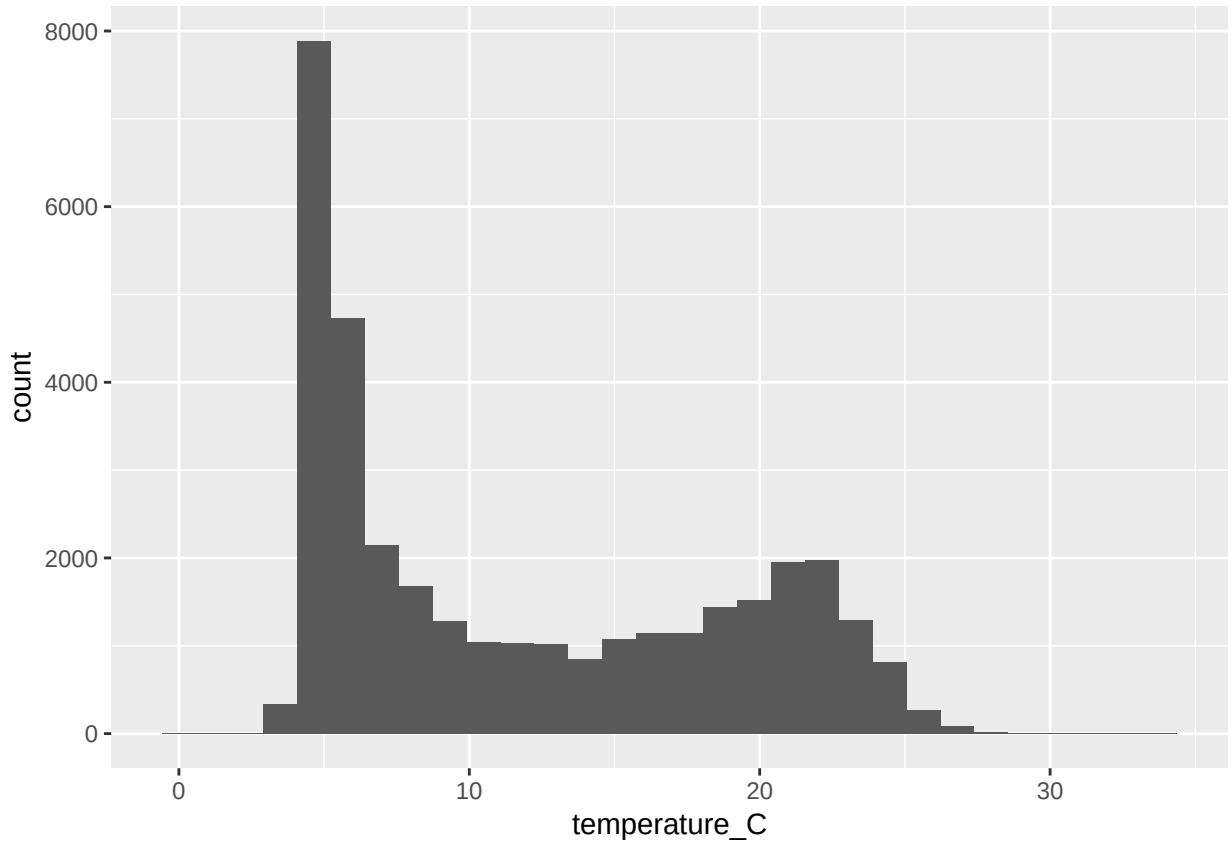



```
ggplot(NTL_LTER_monitoring) +
  geom_histogram(aes(x = temperature_C), bin = 20) # 20 bins
```

```
## Warning: Ignoring unknown parameters: bin
```

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

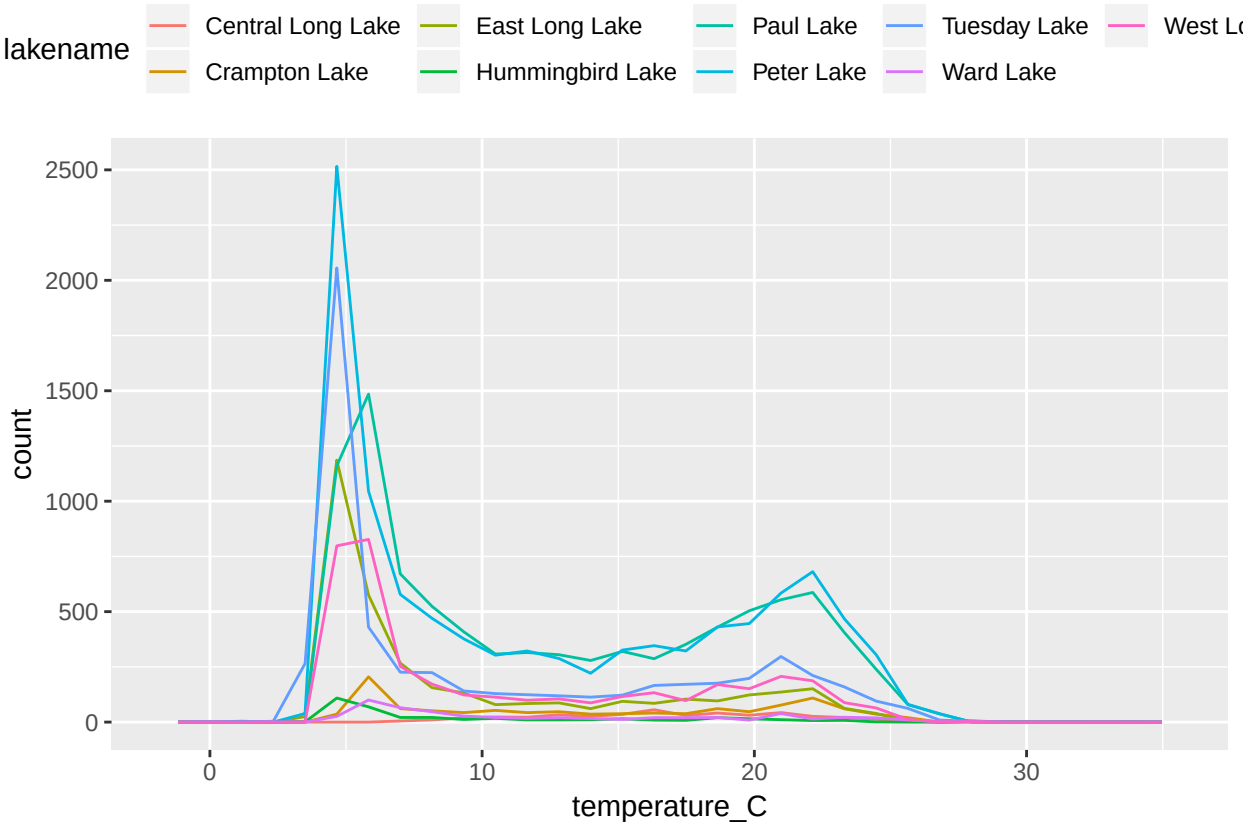
```
## Warning: Removed 3858 rows containing non-finite values (stat_bin).
```




```
# 4
ggplot(NTL_LTER_monitoring) +
  geom_freqpoly(aes(x = temperature_C, color = lakename)) +
  theme(legend.position = "top")
```

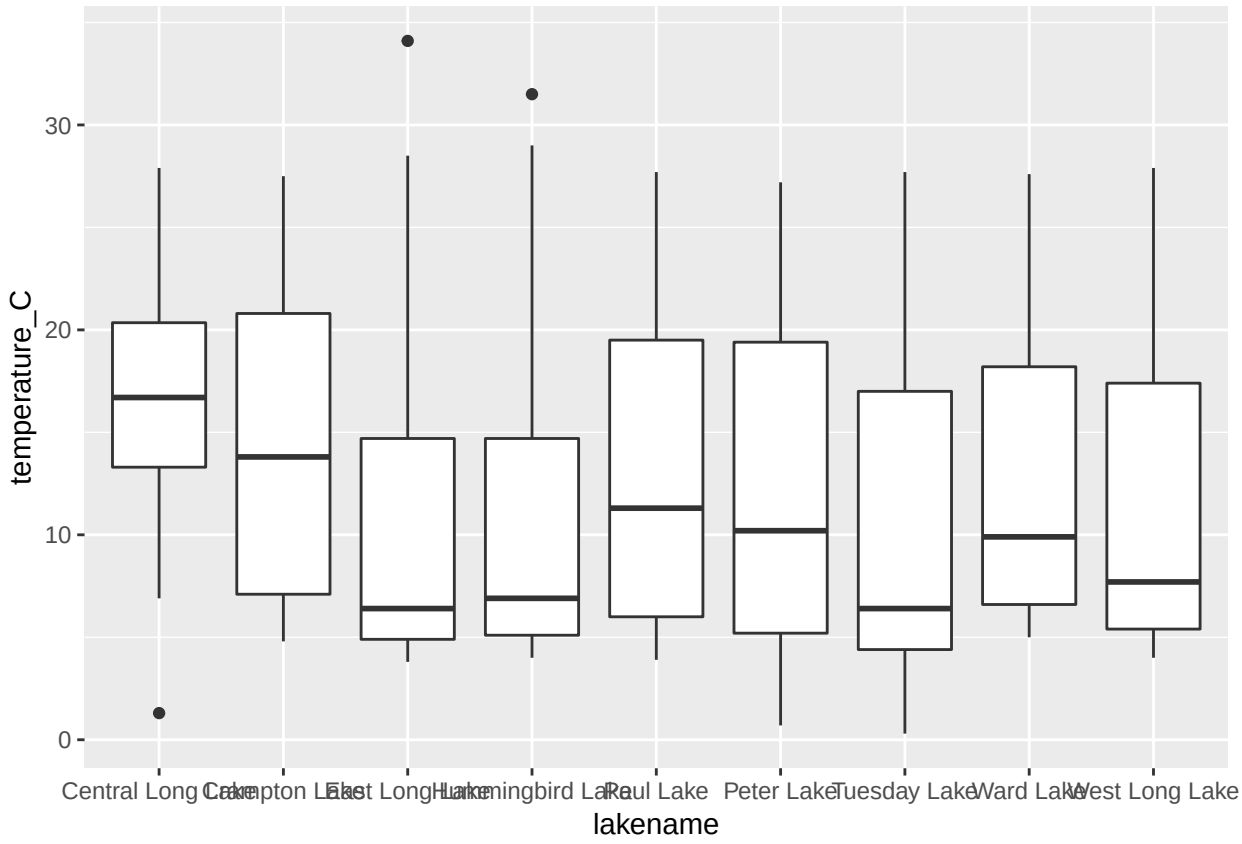
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

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



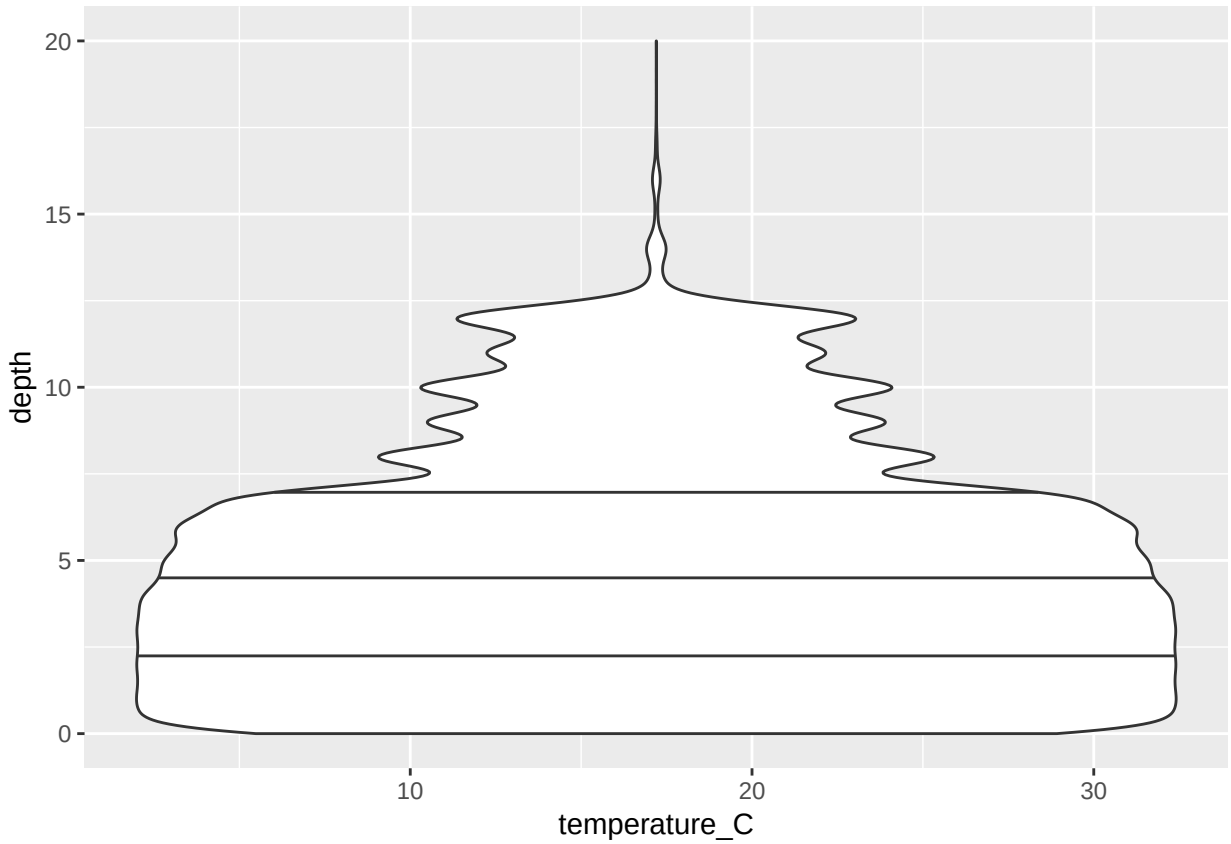

```
# 5
ggplot(NTL_LTER_monitoring) +
  geom_boxplot(aes(x = lakename, y = temperature_C))
```

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



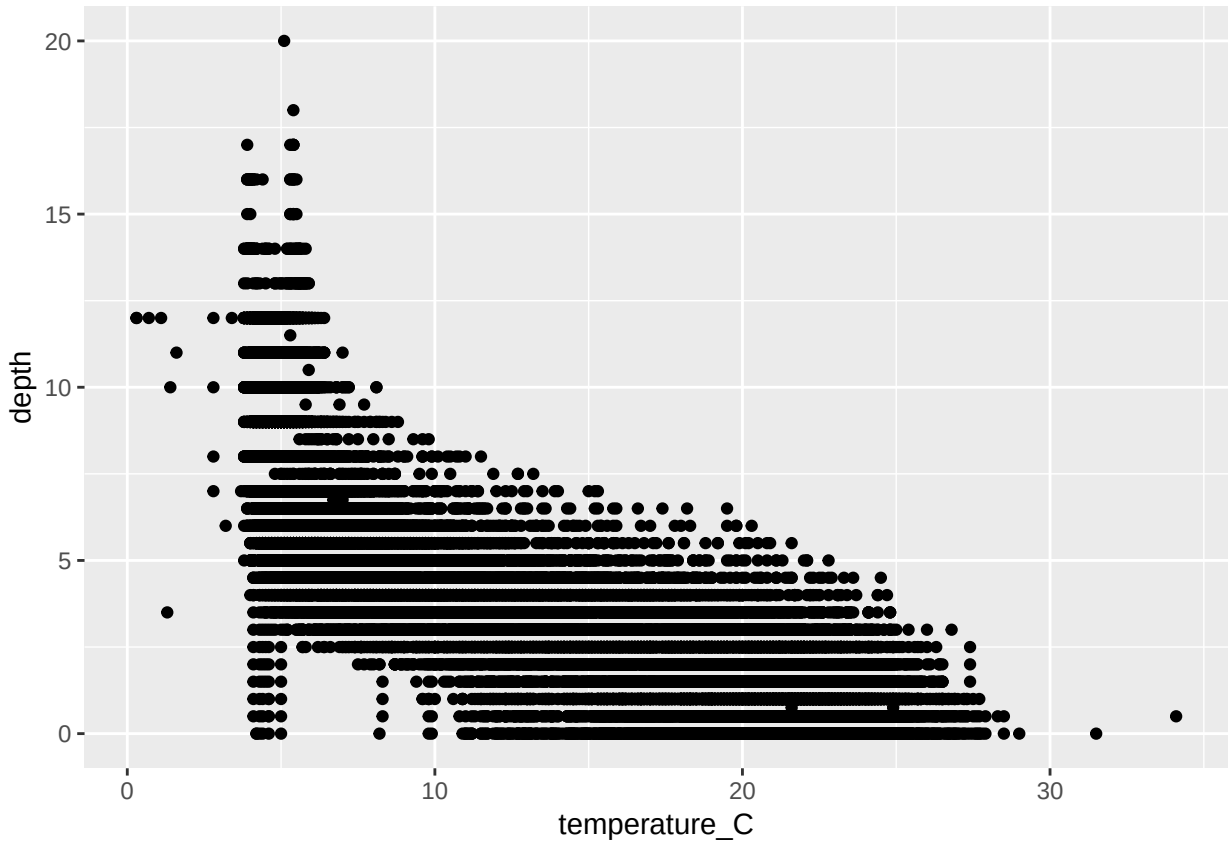

```
# 6
ggplot(NTL_LTER_monitoring) +
  geom_violin(aes(x = temperature_C, y = depth), draw_quantiles = c(0.25, 0.5, 0.75))
```

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




```
# 7
ggplot(NTL_LTER_monitoring) +
  geom_point(aes(x = temperature_C, y = depth))
```

```
## 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: From the basic summaries, I found that Peter Lake has the most records, and Hummingbird Lake has the least records, the overall mean depth of lakes is 4.39m, the overall mean temperature is 11.81 degree with 3858 missing records. From the histogram in step 2, I found that most of the lakes have a temperature around 5 degrees. From the frequency line graph in step 4, I found that most of the lakes share similar temperature profile, which is having a frequency peak at around 5 degree and another small peak at around 22 degree. From the boxplot in step 5, I found that there are 3 outliers within the data. From the violin plot in step 6, I found that the water temperature varies a lot around the surface, but starts to become relatively stable when the depth is over 7m.

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

> ANSWER 1: How does dissolved oxygen change with depth?
>
> ANSWER 2: How does irradiance change with depth, temperature and dissolved oxygen?
>
> ANSWER 3: What are some other factors that affect the temperature/depth gradient? Like lake size or irradiance?