Assignment 3: Data Exploration

Qianyi Xia

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() #check working directory "/Users/xiaqianyi/Documents/current semester/ENV 872/ENV 872/Assignmen
```

```
## [1] "/Users/xiaqianyi/Documents/current semester/ENV 872/ENV 872/Assignments"
```

```
library(tidyverse) #load package
```

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

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.7
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

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

```
LakeChem.data <- read.csv("../Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv") #relative directory
```

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. Explained the background of the data, there are 14 sites, the different characteristic of the lake they study will investigate including depth, dissolved oxygen, temperature, etc. 2.The way the data was sampled. For example the depth intervals of the lake is sampled by the percentage of depth from the surface instead of counting exact number of meters. 3. The carbon data was colected through year 1984 - 2016, while the nutrients data was collected through 1991 - 2016.

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
dim(LakeChem.data)
```

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

```
# 2 class
class(LakeChem.data)
```

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

```
# 3 first 8 rows of the dataset
head(LakeChem.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 the variables lakename, sampledate, depth, and temperature
class(LakeChem.data$lakename)
```

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

```
class(LakeChem.data$sampledate)
```

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

```r
class(LakeChem.data$depth)
```

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

```r
class(LakeChem.data$temperature_C)
```

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

```r
# 5 summary of lakename, depth, and temperature
summary(LakeChem.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(LakeChem.data$depth)
```

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

```r
summary(LakeChem.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
LakeChem.data$sampledate <- as.Date(LakeChem.data$sampledate, format= "%m/%d/%y")
class(LakeChem.data$sampledate)
```

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

```r
head(LakeChem.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, I don't want to remove the NAs. Because there are different types of NAs, while in this dataset, some NAs may because of the below detection value. While for the dissolved oxygen column, the NA may be the value not available. If we remove the NAs arbitrary, the small values may all be removed, there will be big bias with our results. We should think about how to deal with different NAs then decide if we should remove them or assign values to these NAs.

4) Explore your data graphically

Write R commands to display graphs depicting:

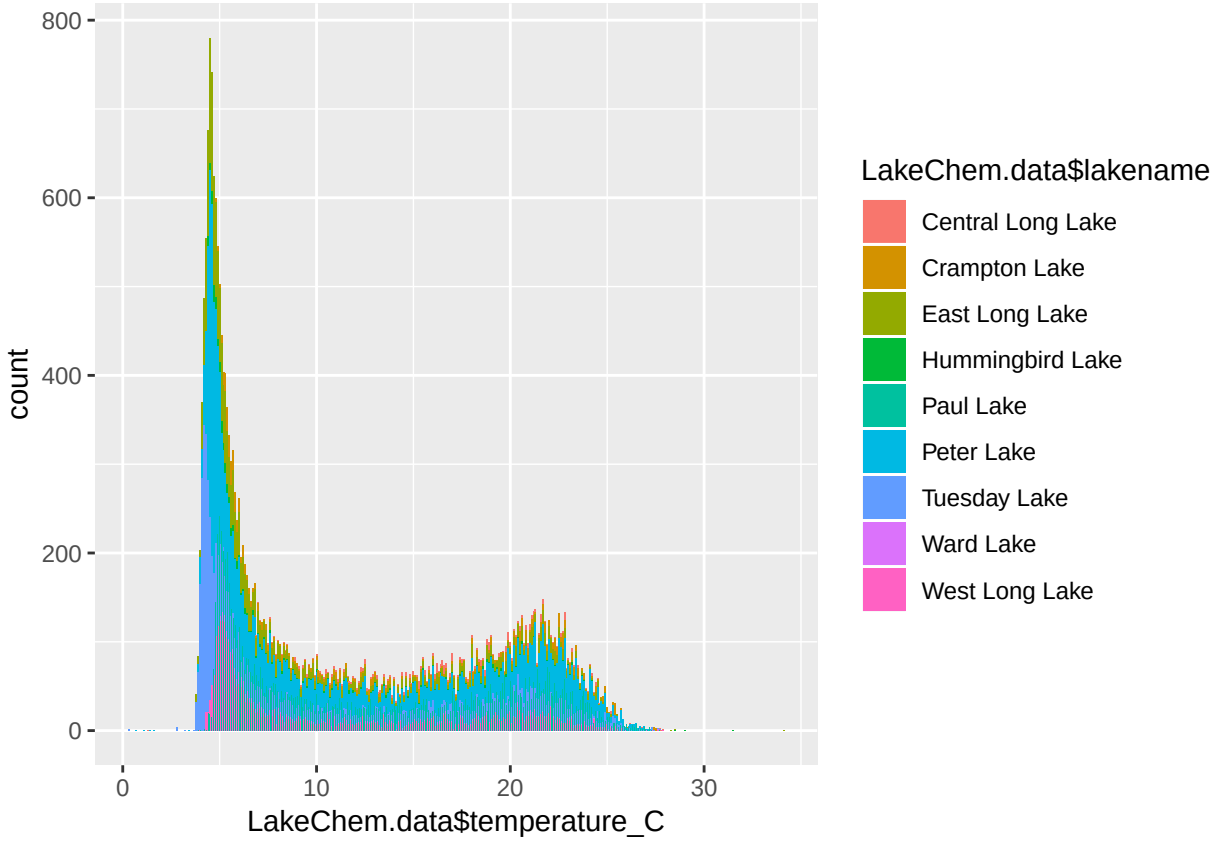
1. Bar chart of temperature counts for each lake
2. Histogram of count distributions of temperature (all temp measurements together)
3. Change histogram from 2 to have a different number or width of bins
4. Frequency polygon of temperature for each lake. Choose different colors for each lake.
5. Boxplot 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 for each lake
ggplot(LakeChem.data, aes(x=LakeChem.data$temperature_C))+
  geom_bar(aes(fill=LakeChem.data$lakename))
```

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

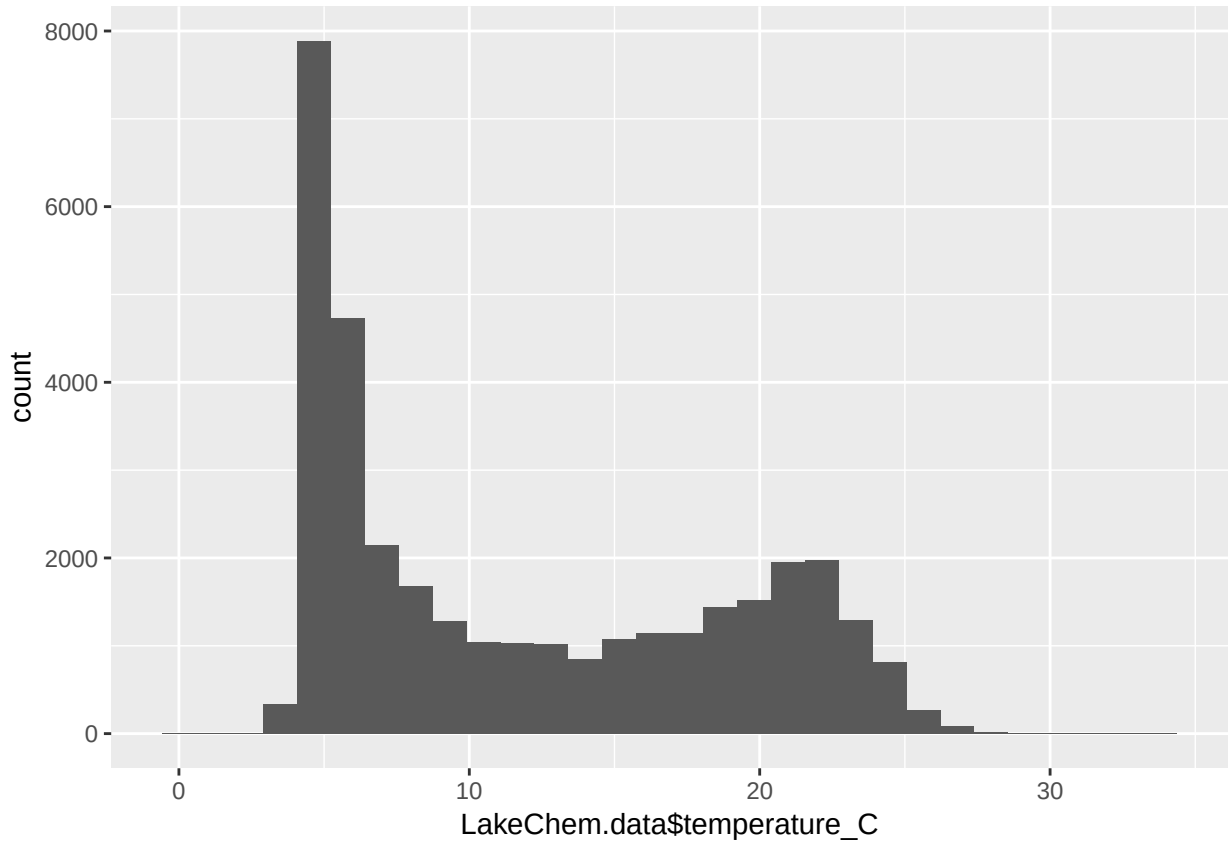
```
## Warning: position_stack requires non-overlapping x intervals
```




```
# 2  Histogram of count distributions of temperature (all temp measurements together)
ggplot(LakeChem.data) +
  geom_histogram(aes(x=LakeChem.data$temperature_C))
```

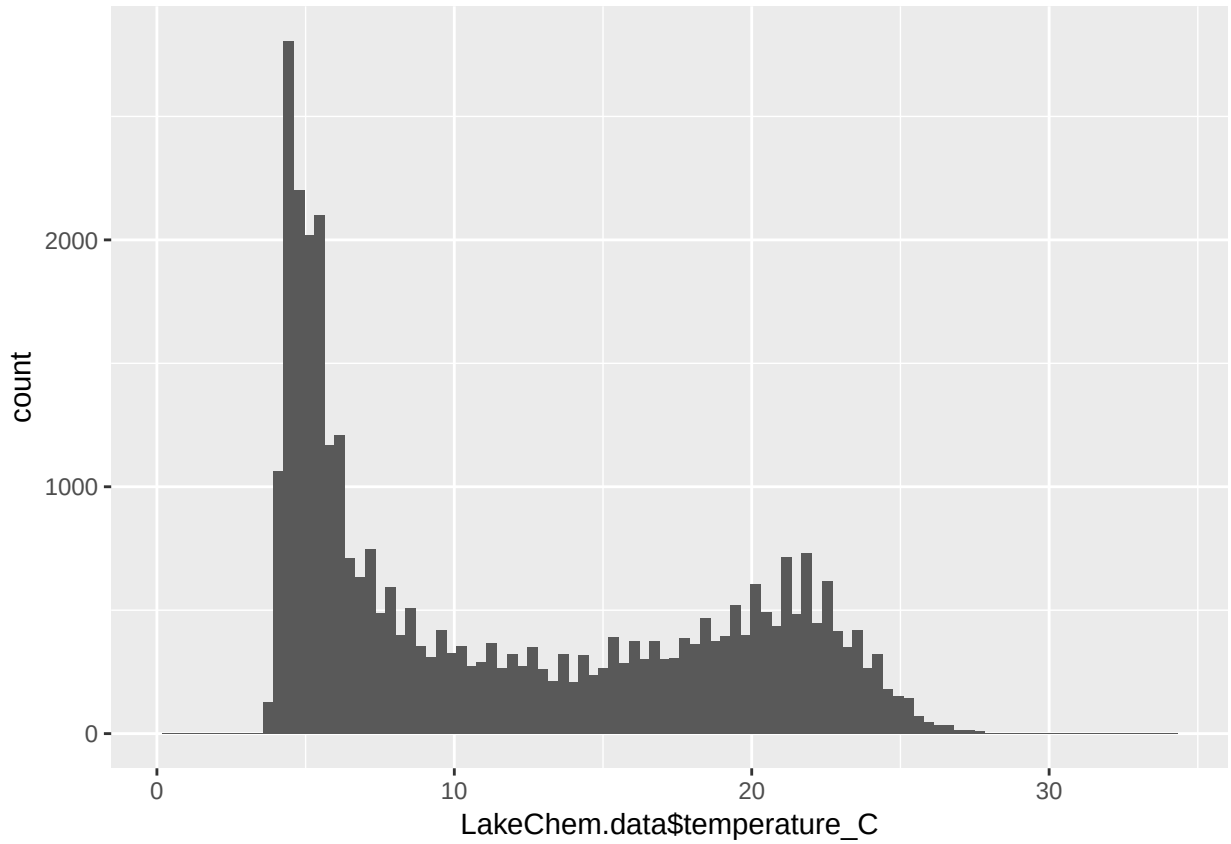
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

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




```
# 3 Change histogram from 2 to have a different number or width of bins
ggplot(LakeChem.data) +
  geom_histogram(aes(x=LakeChem.data$temperature_C), bins = 100)
```

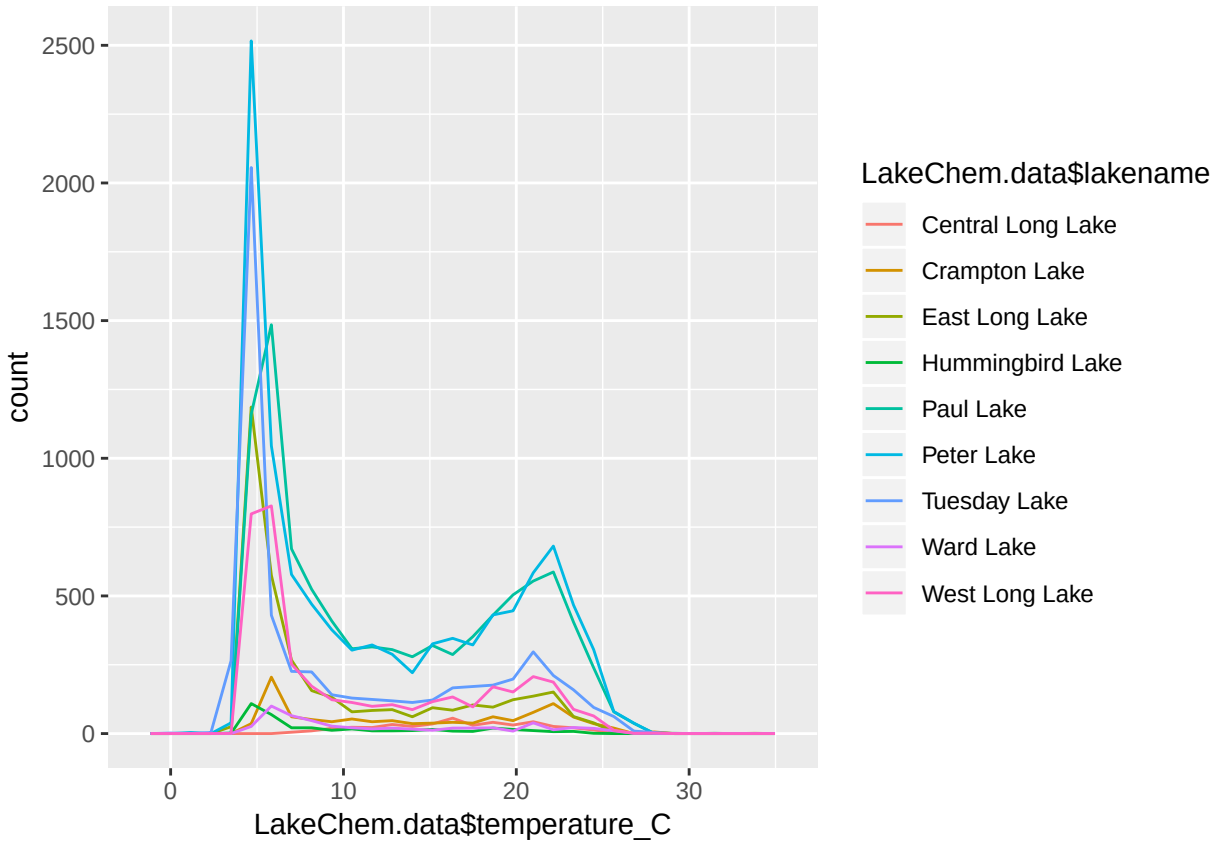
```
## Warning: Removed 3858 rows containing non-finite values (stat_bin).
```




```
# 4 Frequency polygon of temperature for each lake. Choose different colors for each lake
ggplot(LakeChem.data,aes(x = LakeChem.data$temperature_C, fill=LakeChem.data$lakename, colour=LakeChem.d
geom_freqpoly()
```

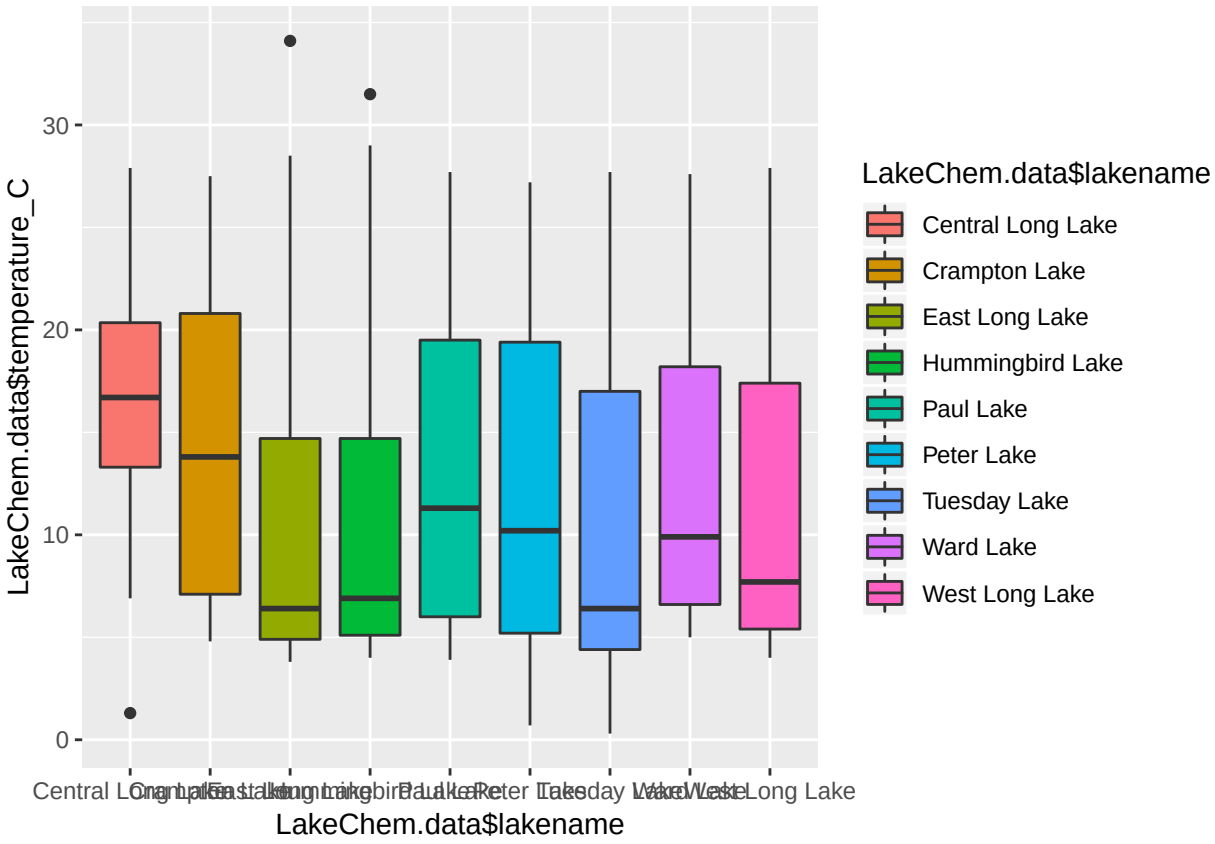
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

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



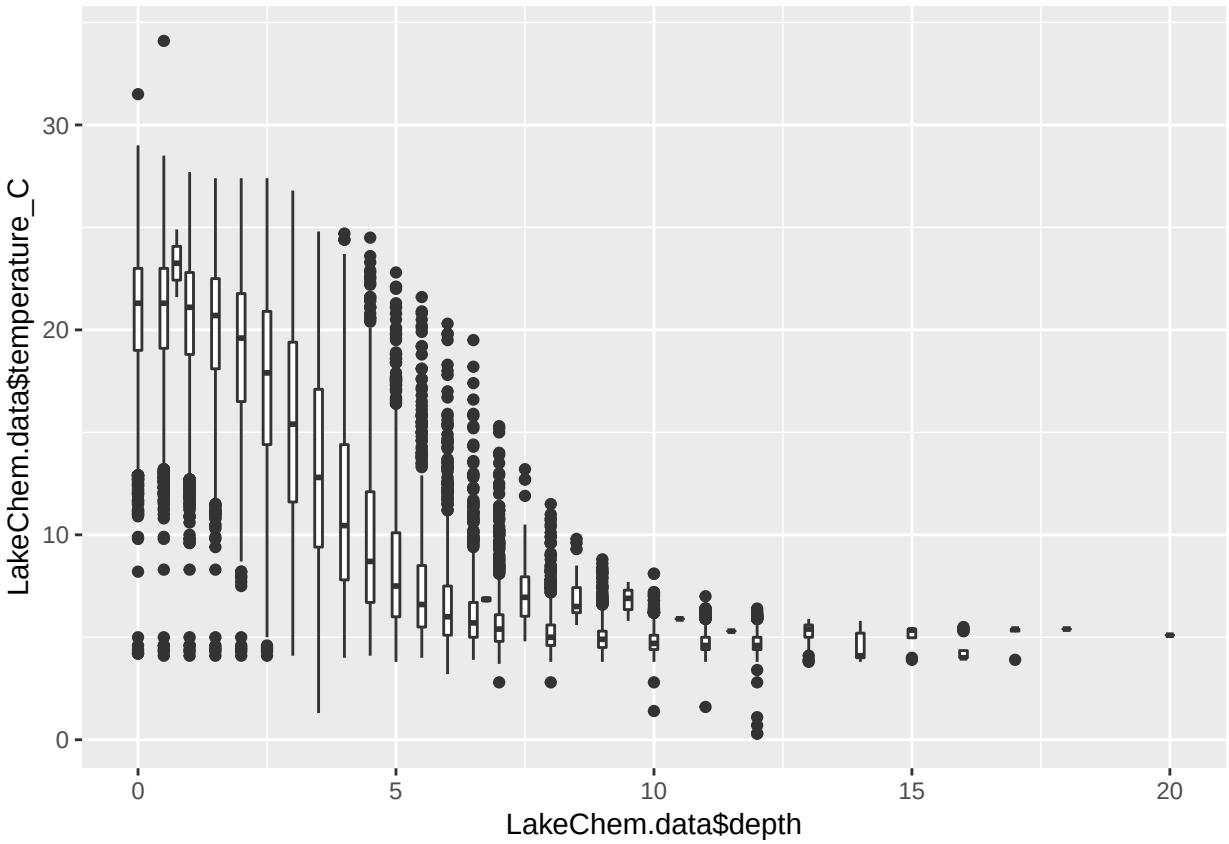

```
# 5 Boxplot of temperature for each lake
ggplot(LakeChem.data)+
  geom_boxplot(aes(x=LakeChem.data$lakename, y=LakeChem.data$temperature_C, fill=LakeChem.data$lakename)
```

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



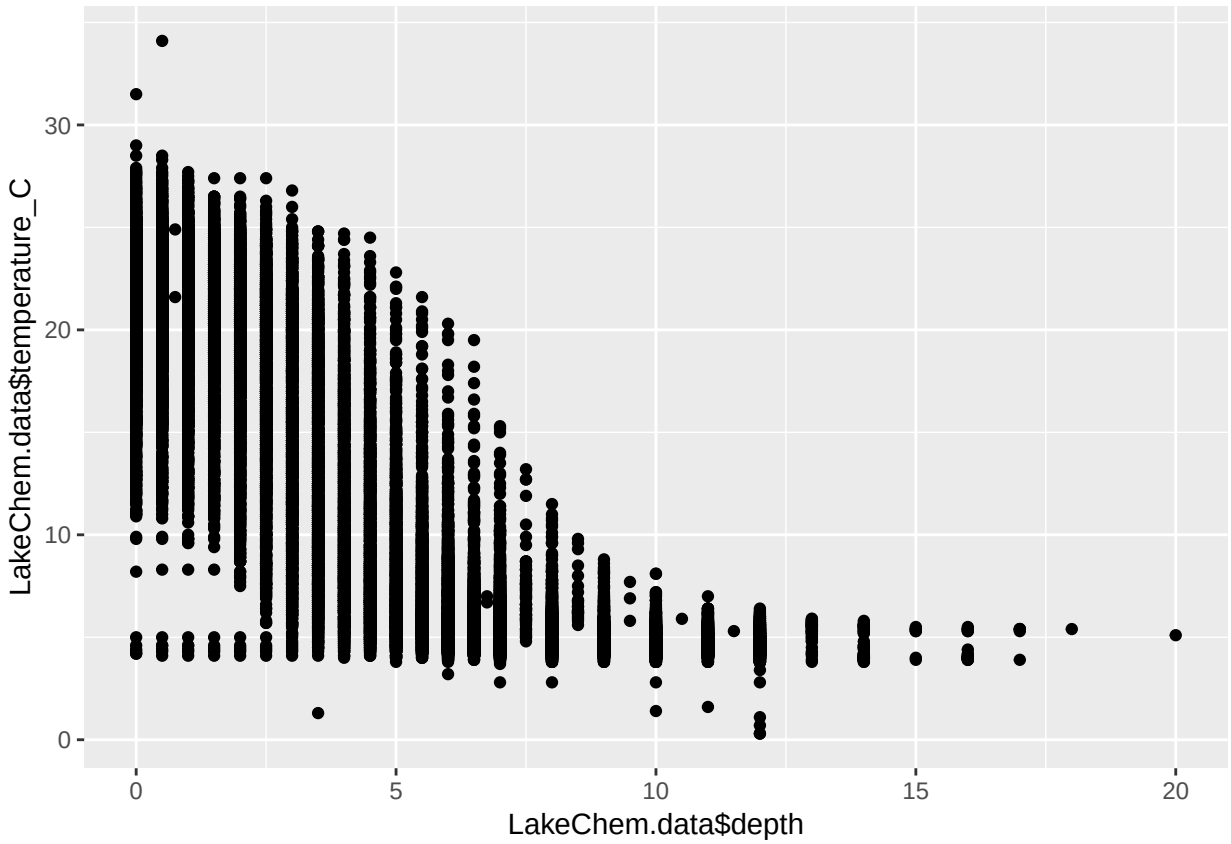

```
# 6 Boxplot of temperature based on depth, with depth divided into 0.25 m increments
ggplot(LakeChem.data)+
  geom_boxplot(aes(x=LakeChem.data$depth, y=LakeChem.data$temperature_C, group=cut_width(LakeChem.data$
```

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




```
# 7 Scatterplot of temperature by depth
ggplot(LakeChem.data)+
  geom_point(aes(x=LakeChem.data$depth, y=LakeChem.data$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: There are 9 Lakes in this dataset. Among all the temperatures measured, most of the temperatures are between 0 to 10 C, and the second most measurements of temperatures are between 20 to 25 C. Peter Lake has the most temperature data collected among the 9 lakes. The Central Long lake has the highest median temperature, while East Long lake and Tuesday Lake have low median temperature among the 9 lakes. And from the figure of temperature by depth, we can get the information that as the depth go deeper in the lake, the temperature will likely to go down and at the same time the temperature range become more stable for there are less surface interactions between water and above environment.

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

> ANSWER 1: How does the dissolved oxygen change as the depth of the Lake go deeper?
>
> ANSWER 2: Is there a relationship between the concentration of dissolved oxygen and surface irradiance?
>
> ANSWER 3: From year 1991 to 2016, how has the temperature of the lakes changed? Does the temperature increased probably due to global warming?