Assignment 3: Data Exploration

Xin Wang

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()
```

```
## [1] "Y:/19spring/872/Environmental_Data_Analytics/Assignments"
```

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.2

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

## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     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

## Warning: package 'ggplot2' was built under R version 3.5.2

## Warning: package 'tibble' was built under R version 3.5.2

## Warning: package 'tidyr' was built under R version 3.5.2

## Warning: package 'readr' was built under R version 3.5.2

## Warning: package 'purrr' was built under R version 3.5.2

## Warning: package 'dplyr' was built under R version 3.5.2

## Warning: package 'stringr' was built under R version 3.5.2

## Warning: package 'forcats' was built under R version 3.5.2
## -- Conflicts ----------------------------------------------------------------------------------- tid
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
NTL.dat <- 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: The source of the data (LTER project), its time span (1984-2016) and collection methodology (location, time, sampling method, etc.).

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.dat)
```

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

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

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

```
# 3
head(NTL.dat,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
attach(NTL.dat)
class(lakename)
```

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

```r
class(sampledate)
```

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

```r
class(depth)
```

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

```r
class(temperature_C)
```

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

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

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

```r
summary(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
```

```r
detach(NTL.dat)
```

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

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

```r
head(NTL.dat$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, because each row represents the measurement at a unique depth for a certain lake on a certain day. There's no repititions, so each row matters, even if it has NAs.

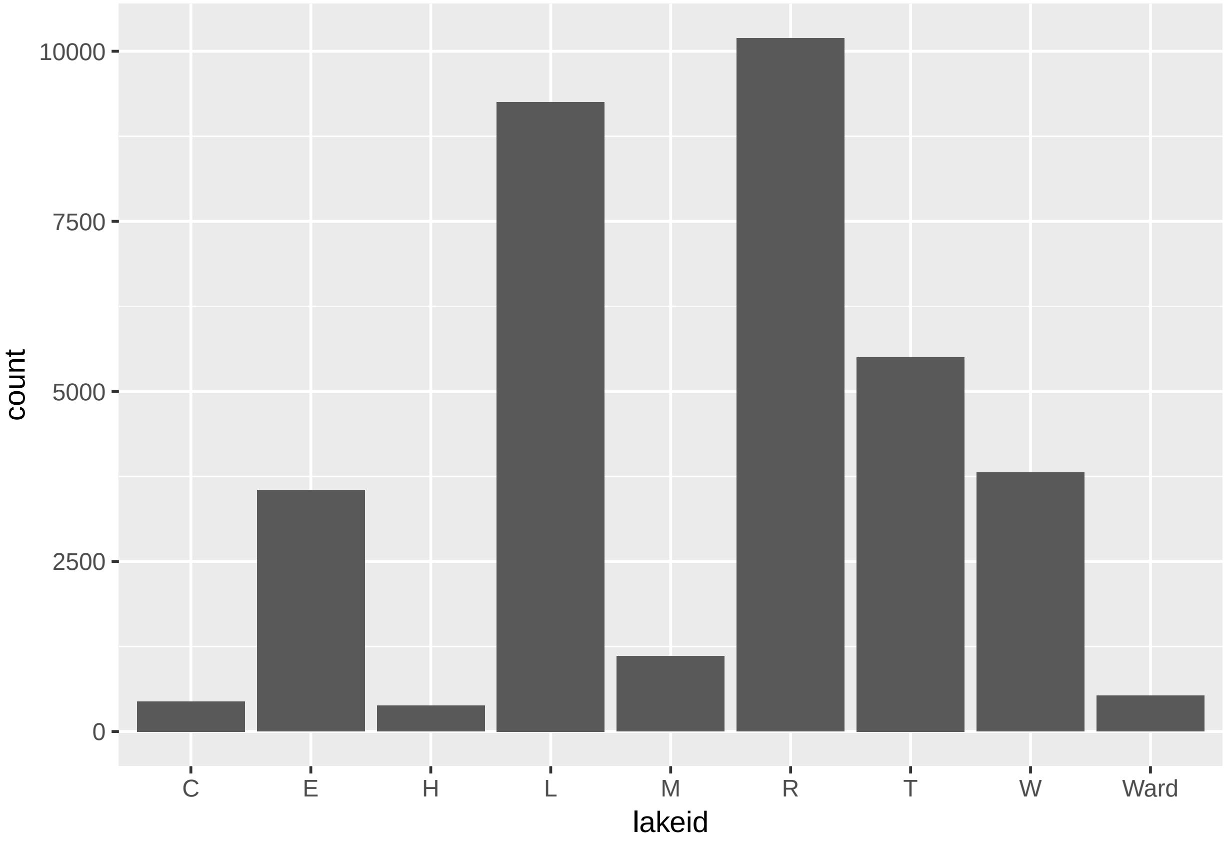
4) Explore your data graphically

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

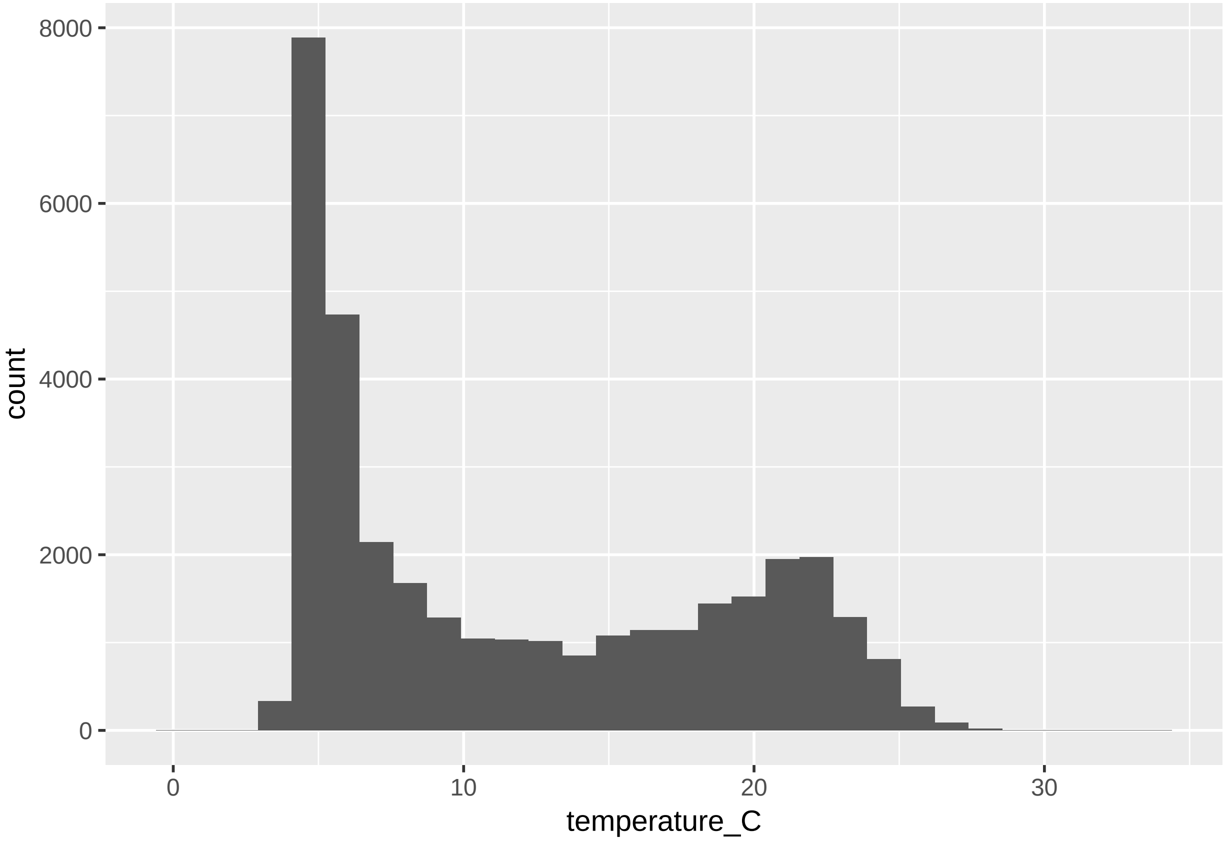
```
# 1
attach(NTL.dat)
temp <- NTL.dat[-which(is.na(temperature_C)),] # generate a dataset with no NAs in the temperature colu
ggplot(temp,aes(x = lakeid)) +
  geom_bar()
```



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

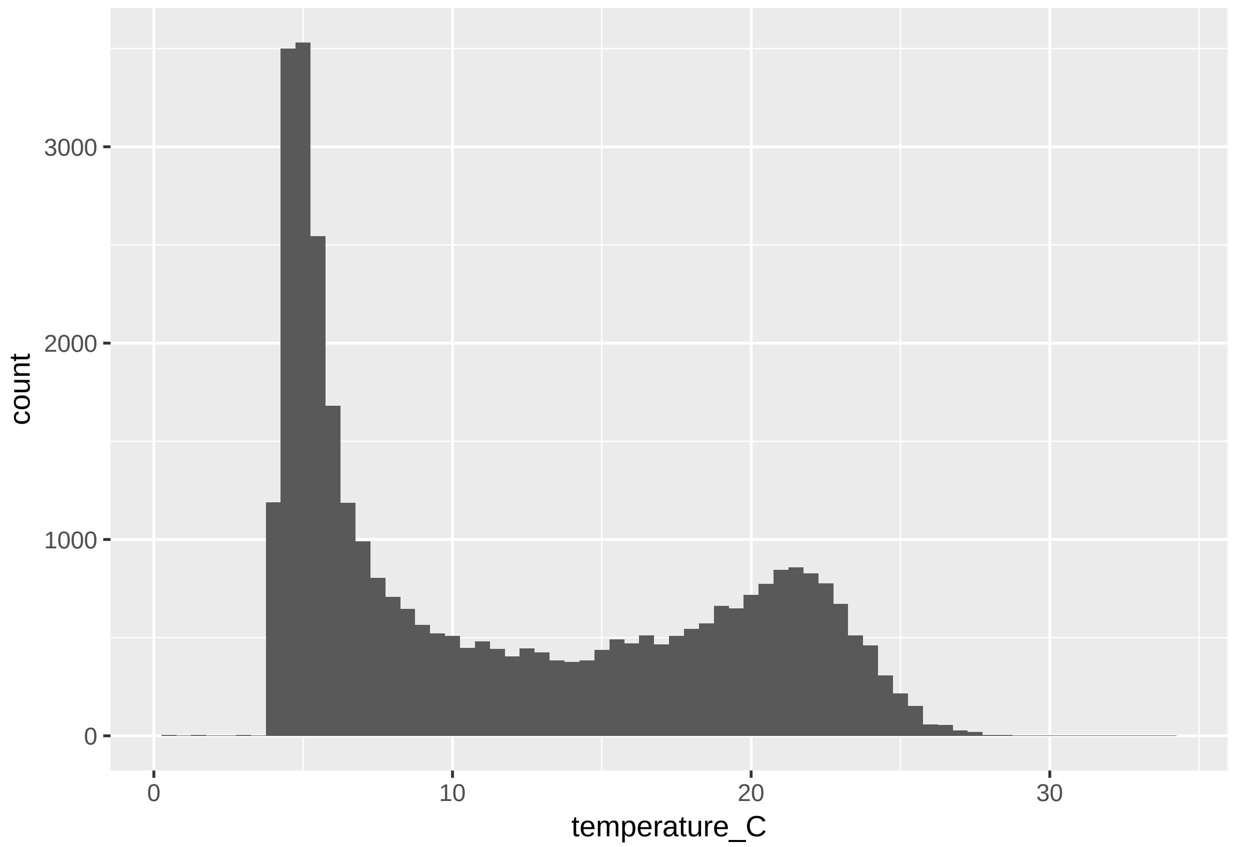
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

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



```
# 3
ggplot(NTL.dat) +
  geom_histogram(aes(x = temperature_C) binwidth = 0.5)
```

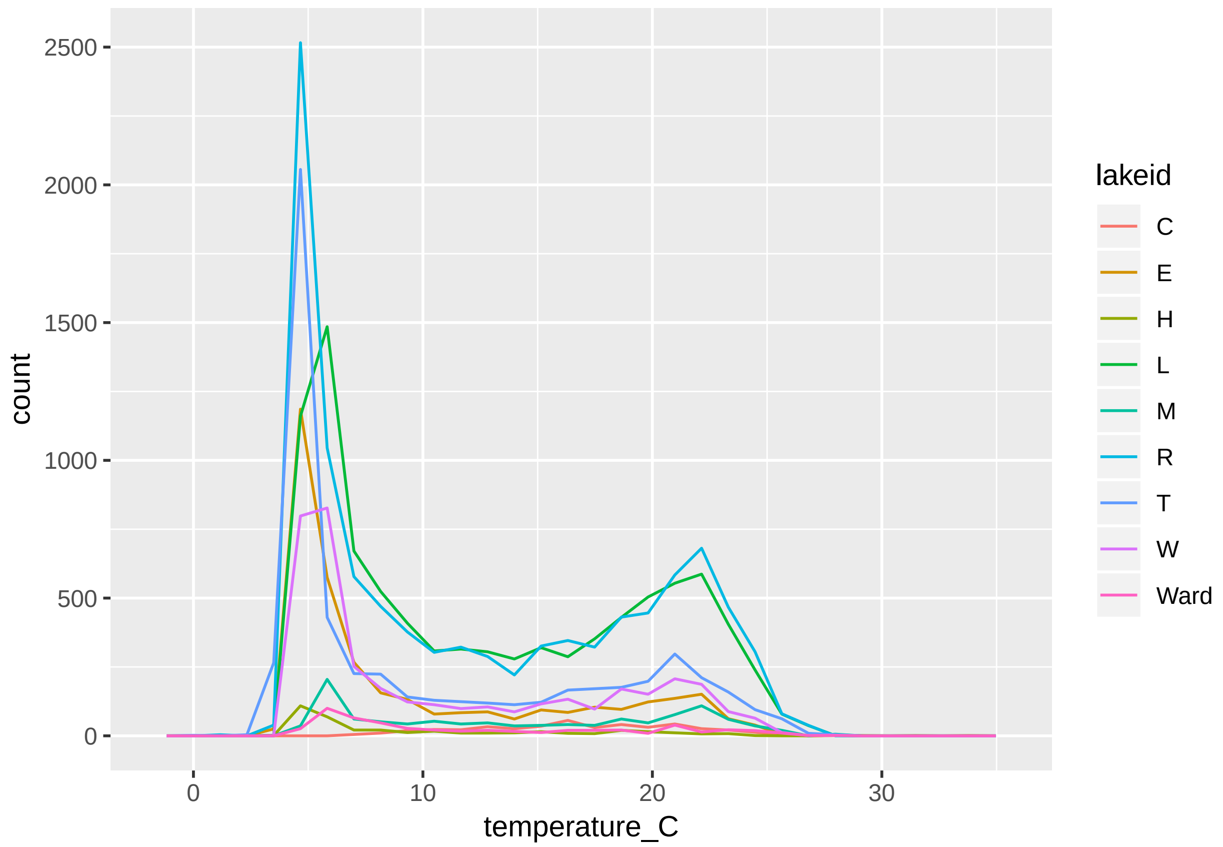
```
## Warning: Removed 3858 rows containing non-finite values (stat_bin).
```



```r
# 4
ggplot(NTL.dat) +
  geom_freqpoly(aes(x = temperature_C, color = lakeid))
```

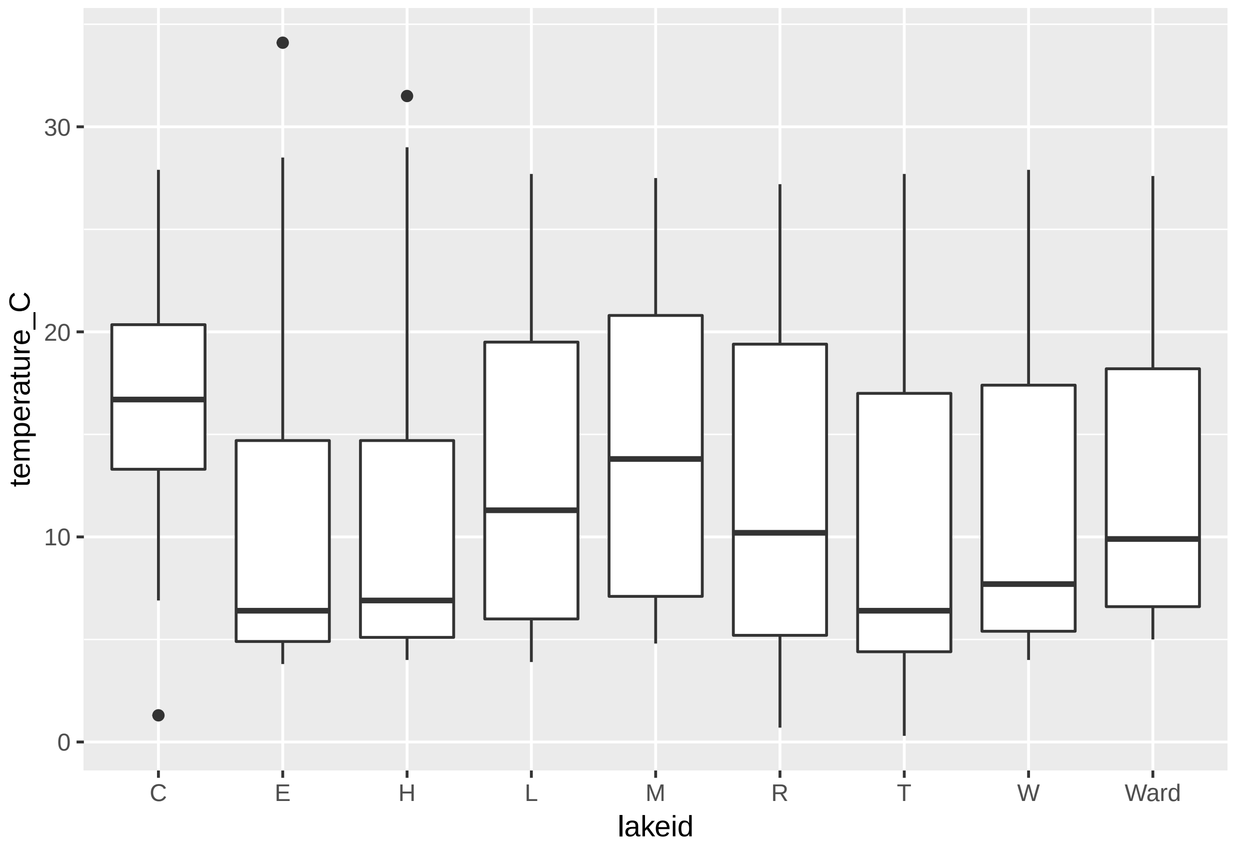
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

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



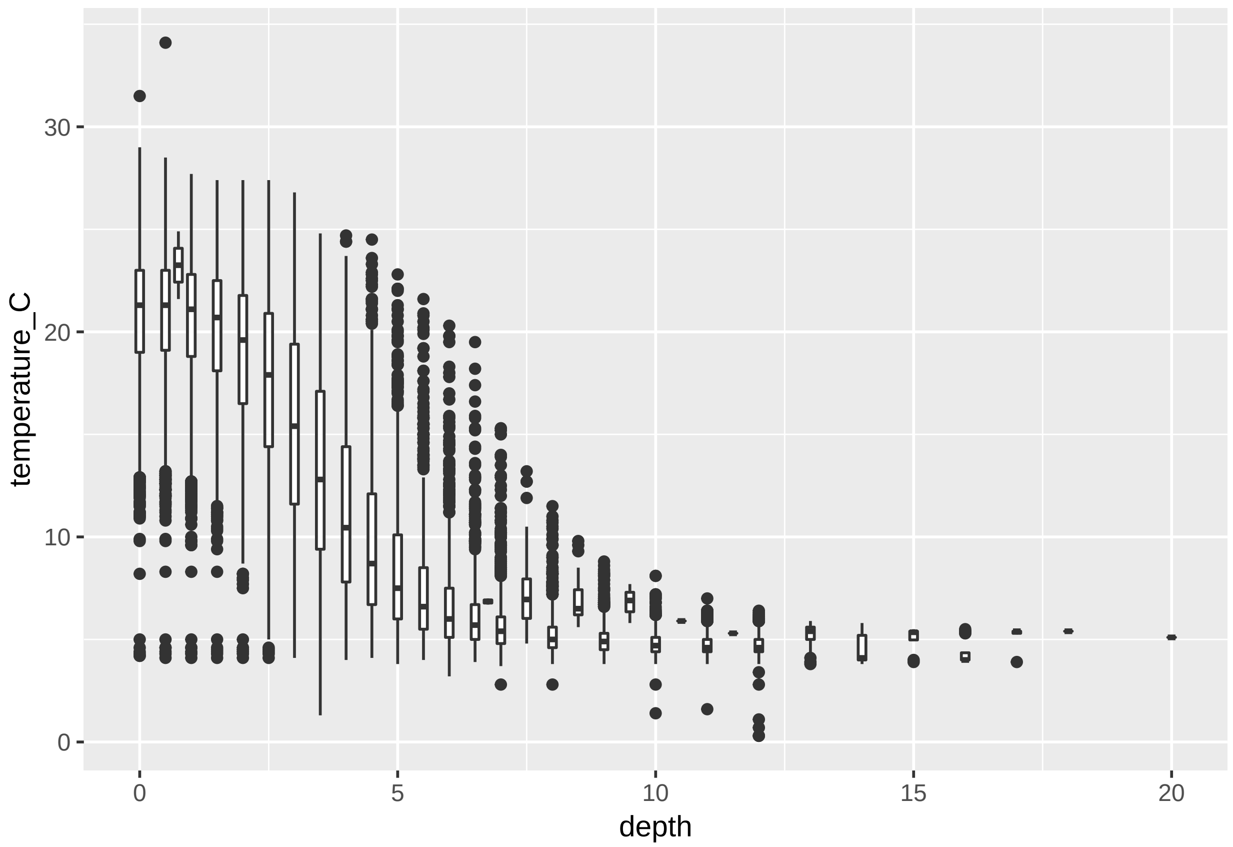
```
# 5
ggplot(NTL.dat) +
  geom_boxplot(aes(x = lakeid,y = temperature_C))
```

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



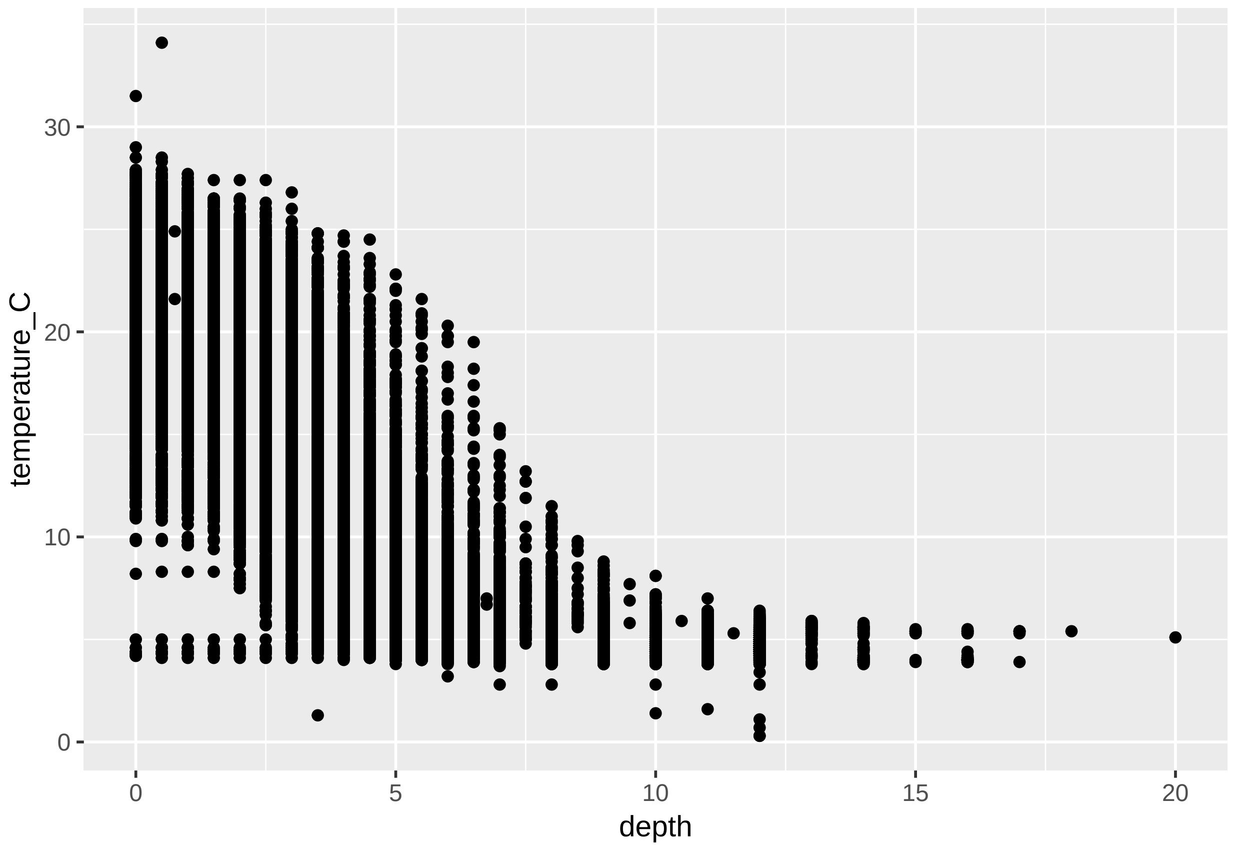
```
# 6
ggplot(NTL.dat) +
  geom_boxplot(aes(x = depth, y = temperature_C, group = cut_width(depth, 0.25)))
```

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



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

```
## Warning: Removed 3858 rows containing missing values (geom_point).
```



```
detach(NTL.dat)
```

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 totally 9 lakes investigated for the project, with Paul Lake and Peter Lake the most samples. Sampling depth ranges from 0 to 20m, but most samples are collected in water shallower than 10m. The frequency distribution of temperature among the samples has two peaks, one around 5 degree celcius and the other around 22 degree celcius. Generally, water temperature decreases as depth increases, but remains steady in water deeper than 10m.

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: For the physical and chemical variables, is there significant difference among different lakes?

> ANSWER 3: How do the physical and chemical variables of the lakes change with time?