Assignment 3: Data Exploration

Xin Zhang

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()
```

```
## [1] "C:/Users/Xin Zhang/Desktop/EDA/Assignments"
```

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

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

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

```
#upload the dataset
TempLakes.monitor.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) This dataset contains data from studies on several lakes in the North Temperate Lakes District in Wisconsin, USA from 1984 to 2016 and it includes the following data contents: names, sample depths, date, temperature,physical and chemical limnology indicators and etc. (2) This dataset comes from the North Temperate Lakes Long Term Ecological Research website https://lter.limnology.wisc.edu/data.Four selctions were made to select the data we need. (3) Data were accessed 2018-12-06.

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 Demnsions
dim(TempLakes.monitor.data)
```

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

```
# 2 class
class(TempLakes.monitor.data)
```

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

```
# 3 first 8 rows
head(TempLakes.monitor.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 lakename, sampledate, depth and temperature
class(TempLakes.monitor.data$lakename)
```

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

```
class(TempLakes.monitor.data$sampledate)
```

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

```r
class(TempLakes.monitor.data$depth)
```

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

```r
class(TempLakes.monitor.data$temperature_C)
```

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

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

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

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

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

```r
head(TempLakes.monitor.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?

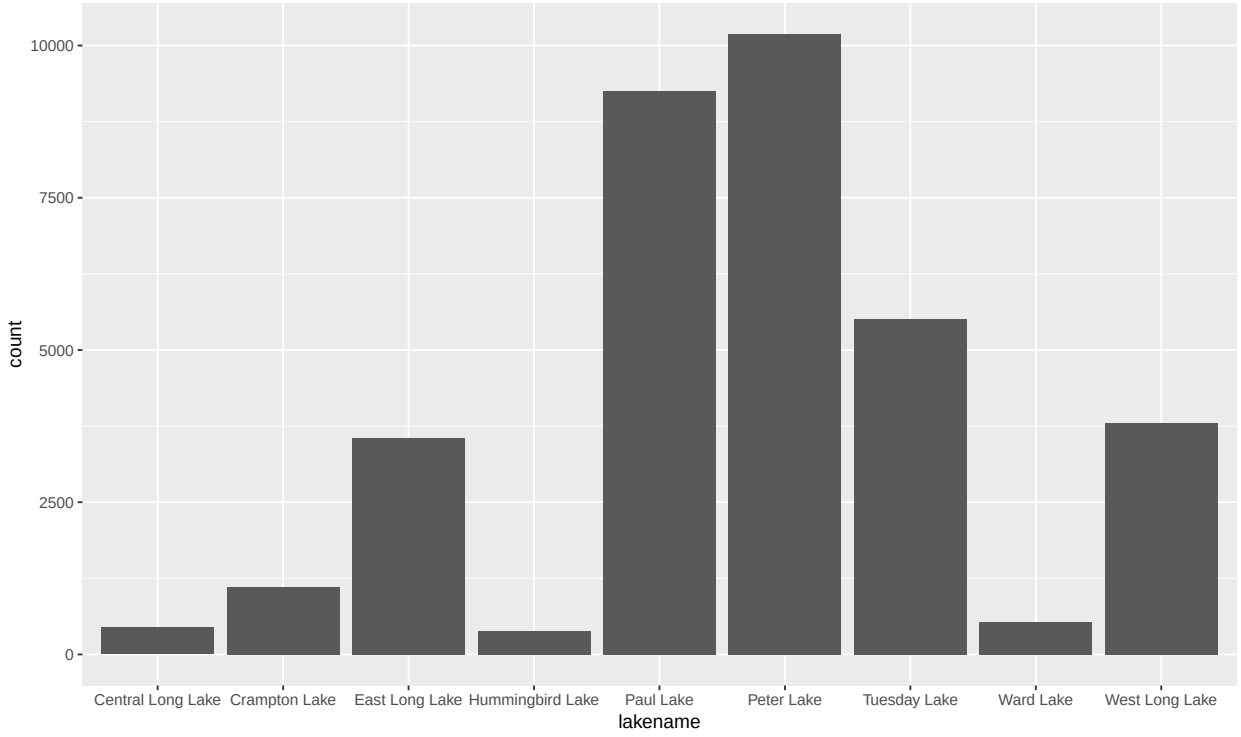
> ANSWER: There is a column named 'comments' and all values in it are NA, if i use na.omit(), it will delete all the rows. And, when using ggplot for graphs, it will automatically remove the NAs in the variables we use. Therefore, I will not remove the NAs. However, if there is analysis that needs me to remove some specific NAs, I will work on it. At this point, i would say, I will not use na.omit to delete all my data.

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

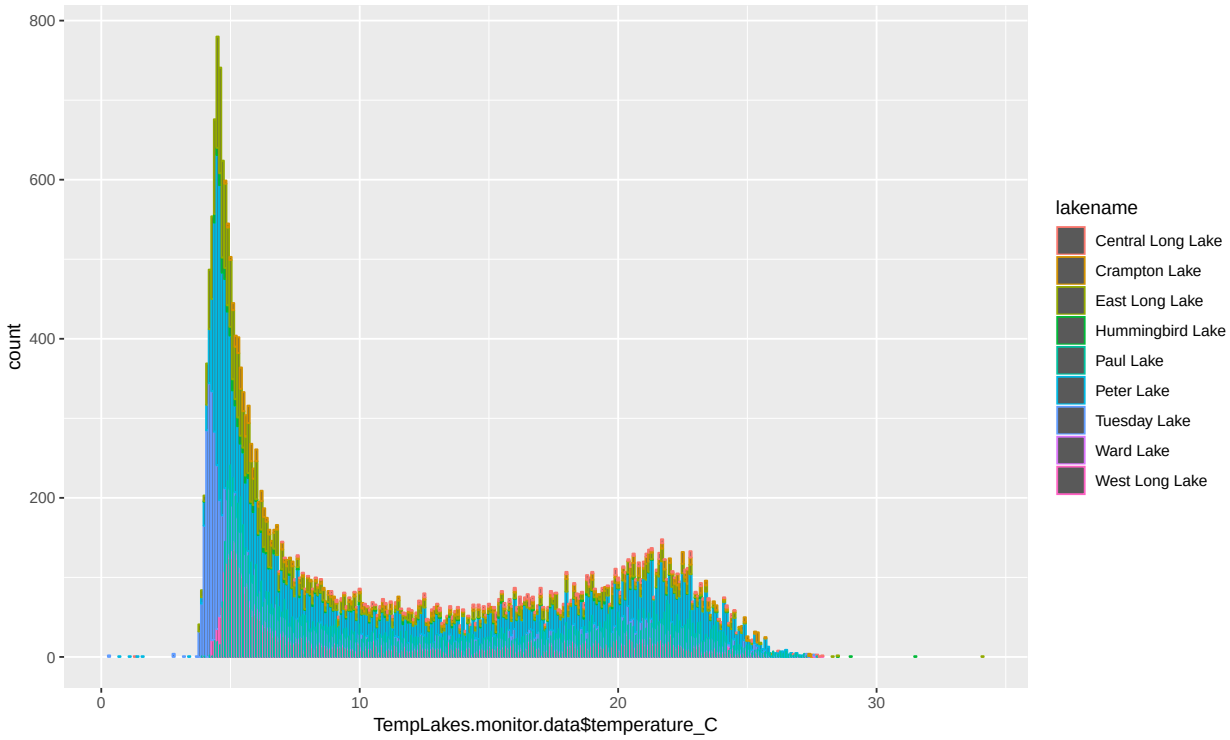
```r
# 1
#if 'temperature counts' means the total number of temperature records for each lake
ggplot(TempLakes.monitor.data[!is.na(TempLakes.monitor.data$temperature_C), ], aes(x = lakename)) +
  geom_bar()
```




```r
#if 'temperature counts' means the number of each temperatue at each lake
ggplot(TempLakes.monitor.data, aes(x = TempLakes.monitor.data$temperature_C, color = lakename )) +
  geom_bar()
```

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

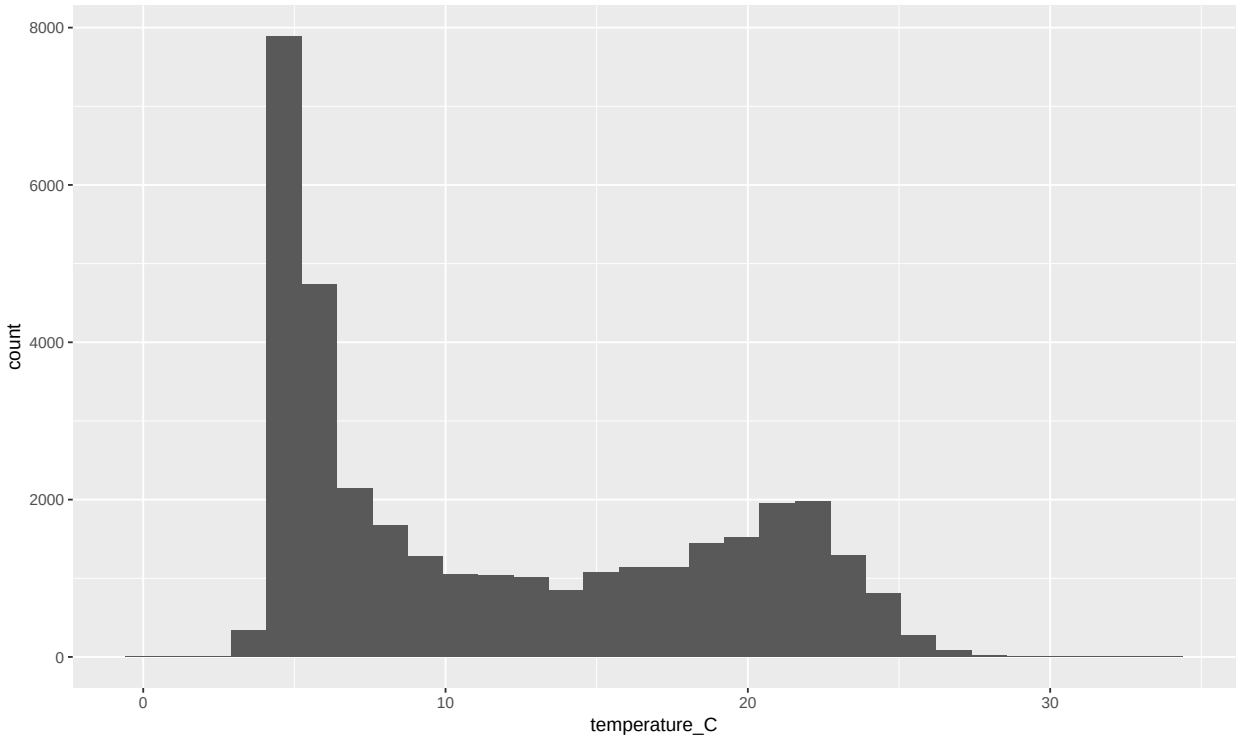
```
## Warning: position_stack requires non-overlapping x intervals
```




```
# 2
ggplot(TempLakes.monitor.data) +
  geom_histogram(aes(x = temperature_C))
```

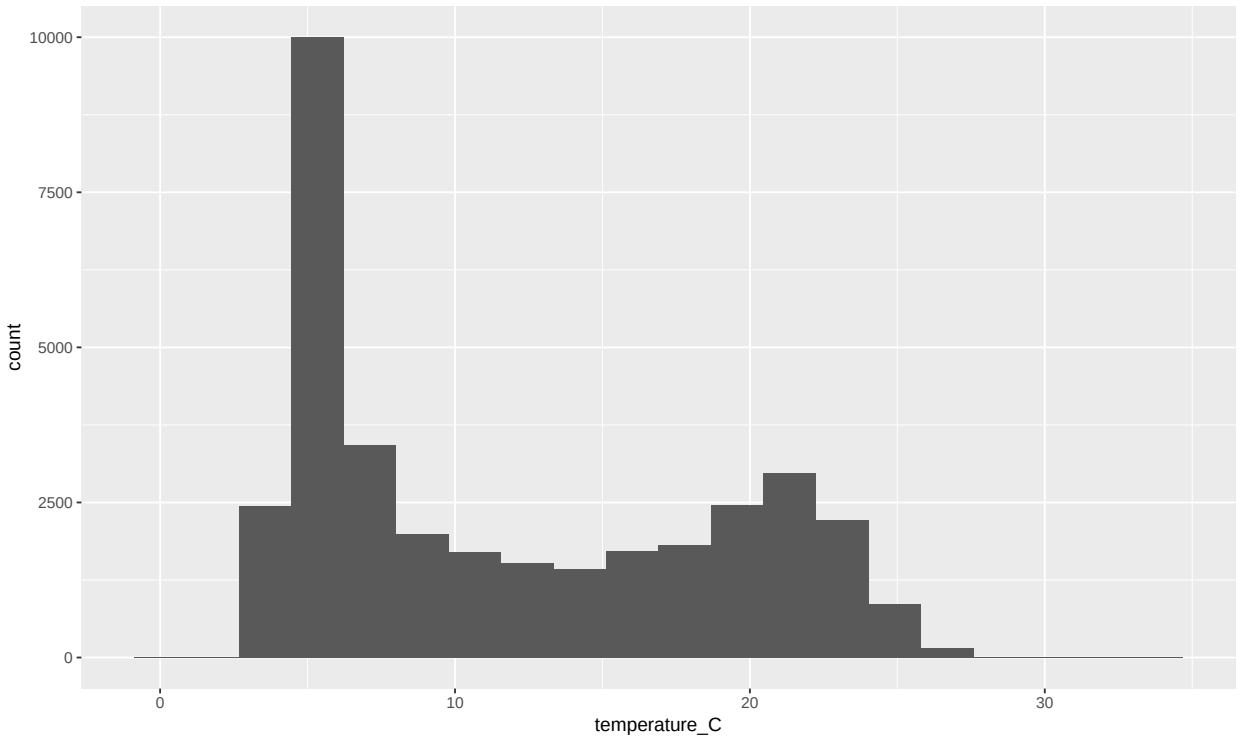
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

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



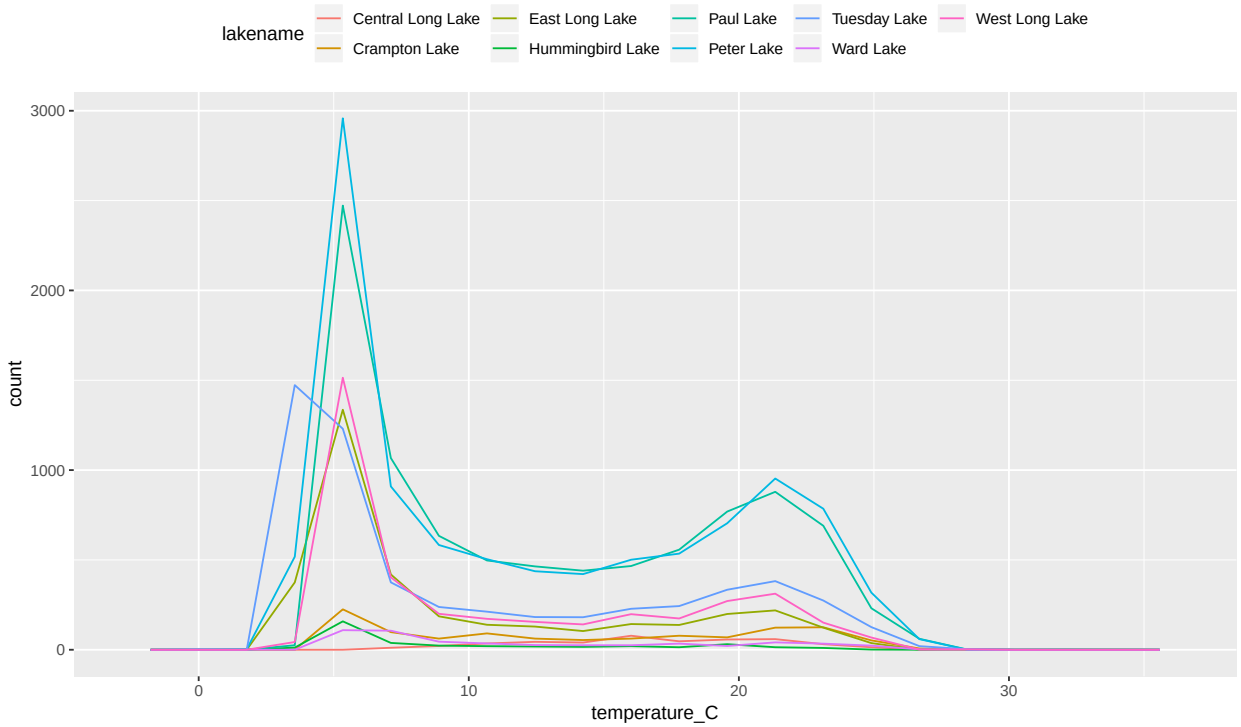

```
# 3
ggplot(TempLakes.monitor.data) +
  geom_histogram(aes(x = temperature_C),bins=20)
```

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



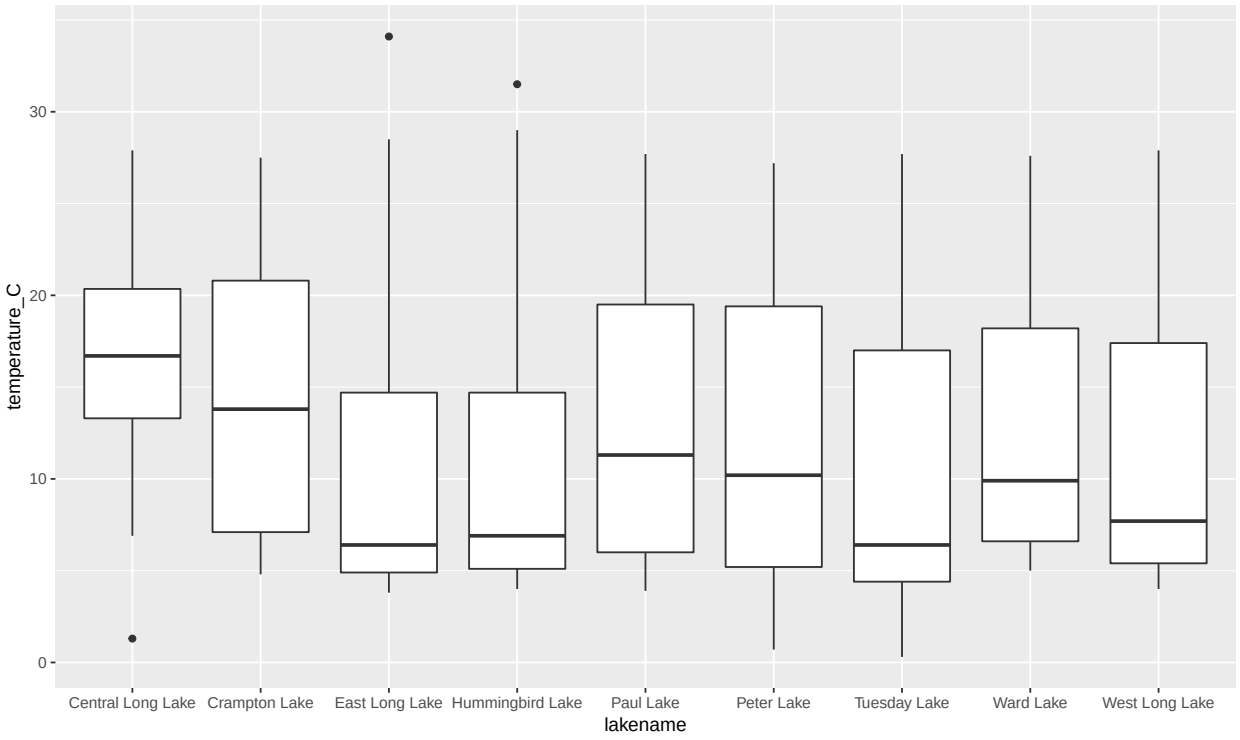
```
# 4

ggplot(TempLakes.monitor.data) +
  geom_freqpoly(aes(x = temperature_C, color = lakename), bins = 20) + theme(legend.position = "top")
```

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



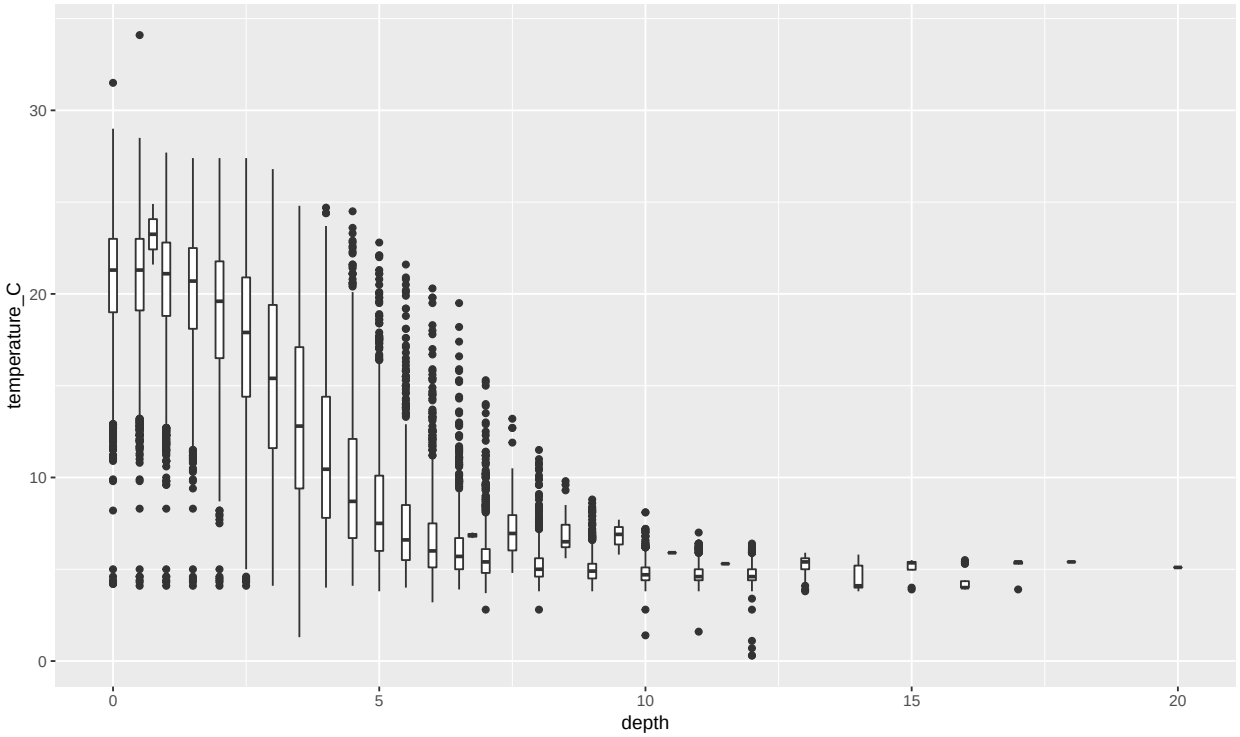

```
# 5
ggplot(TempLakes.monitor.data) +
  geom_boxplot(aes(x = lakename, y = temperature_C))
```

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



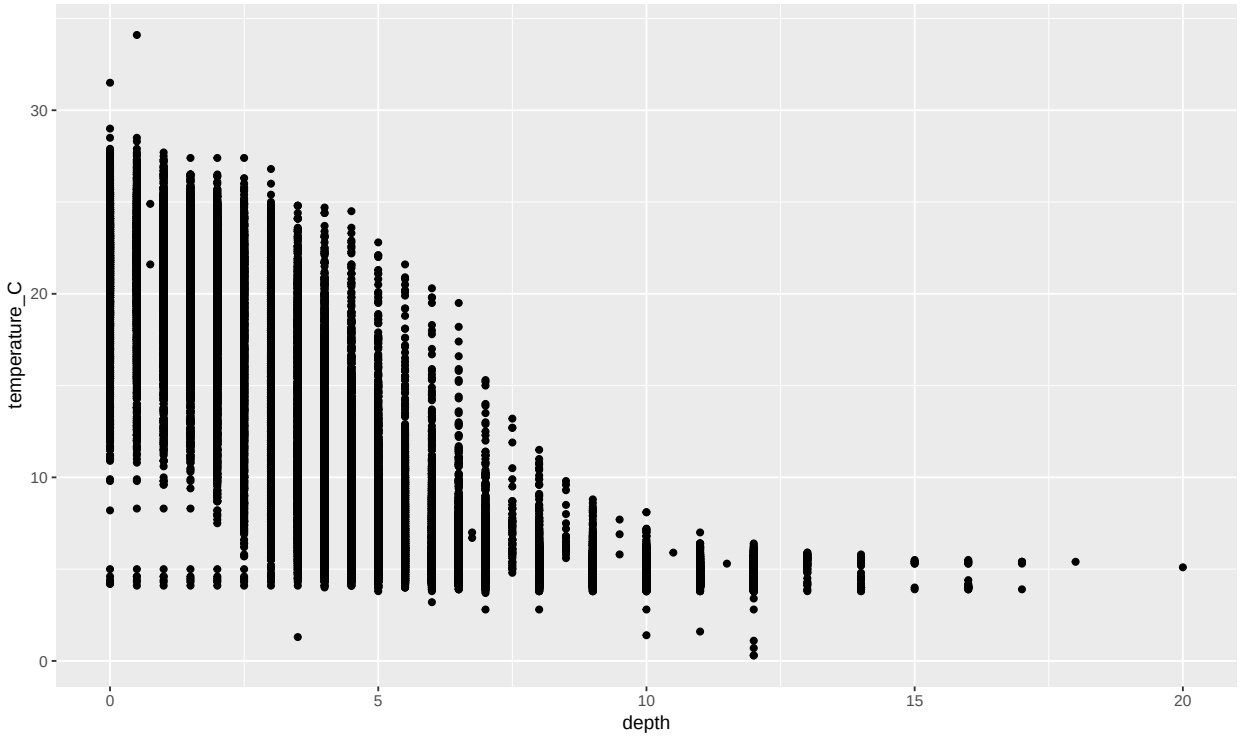

```
# 6
ggplot(TempLakes.monitor.data) +
  geom_boxplot(aes(x = depth, y = temperature_C, group = cut_width(depth, 0.25)))
```

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



```r
# 7
ggplot(TempLakes.monitor.data) +
  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: There are 9 temperate lakes in total in Wisconsin from 1984 to 2016 and Peter Lake has the most records. The lake depth ranged from 0 to 20 and the mean depth is 4.39; the temperature ranged from 0.3 to 34.1, the mean temperature is 11.81 and there was 3858 records of missing temperature. The most frequent showed temperature was around 5 Celsius, and temperature distribution showed two peaks, one at 5 and another at 22.These nine temperate lakes had similar temperature distribution pattern, and each lake's most temeprature records ranged around 5 to 22, but there were some outliers at each lake's records. When sample depth of the lake increased, the temperature decreased, and when the depth increased to more than 10, the temperature droped more slowly and reamained around 5 Celsius.

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

> ANSWER 1: Is there any relation between temperature and dissolved Oxygen, when temperature increases, would dissolved Oxygen change in certain trend?
>
> ANSWER 2: Is there any relation between depth and dissolved Oxygen, when depth increases, would dissolved Oxygen change in certain trend?
>
> ANSWER 3: Is there any temperature change trend in time series of 1984 to 2016 in each lake?