Assignment 3: Data Exploration

Jiaqi Li

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] "/Users/ljq/Desktop/Blue Devils/Data Analysis/ENV872_02/Assignments"
```

```
setwd("/Users/ljq/Desktop/Blue Devils/Data Analysis/ENV872_02")
# Load package
library(tidyverse)
```

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

## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.7
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0

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

```
#Import Data
lake <- 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: Three salient pieces of information in the README file are database information, data content information, and naming conventions and file formats. Database information describes where, how, and when the data was collected. Data content information contains definitions and collection methods of each variable from the NTL-LTER database website, including carbon, nutrients, and physical and chemical limnology. Naming conventions and file formats contains the rules of naming files.

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

```r
# 1 Dimension of the dataset
dim(lake)
```

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

```r
# 2 Class of the dataset
class(lake)
```

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

```r
# 3 First 8 rows of the dataset
head(lake, 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>
```

```r
# 4 Class of the variables
class(lake$lakename); class(lake$sampledate); class(lake$depth); class(lake$temperature_C)
```

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

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

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

```
# 5 Summary of the variables
summary(lake$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
```

```
summary(lake$depth)
```

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

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

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

```
head(lake$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, the reason that NAs should not be removed is that NAs have meanings that some of the data was not collected or not availble in this dataset. It may cause inaccuracy in the data analysis if NAs are removed.

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(lake, aes(x = temperature_C)) +
  geom_bar(aes(fill = lakename, binwidth=0.5))
```

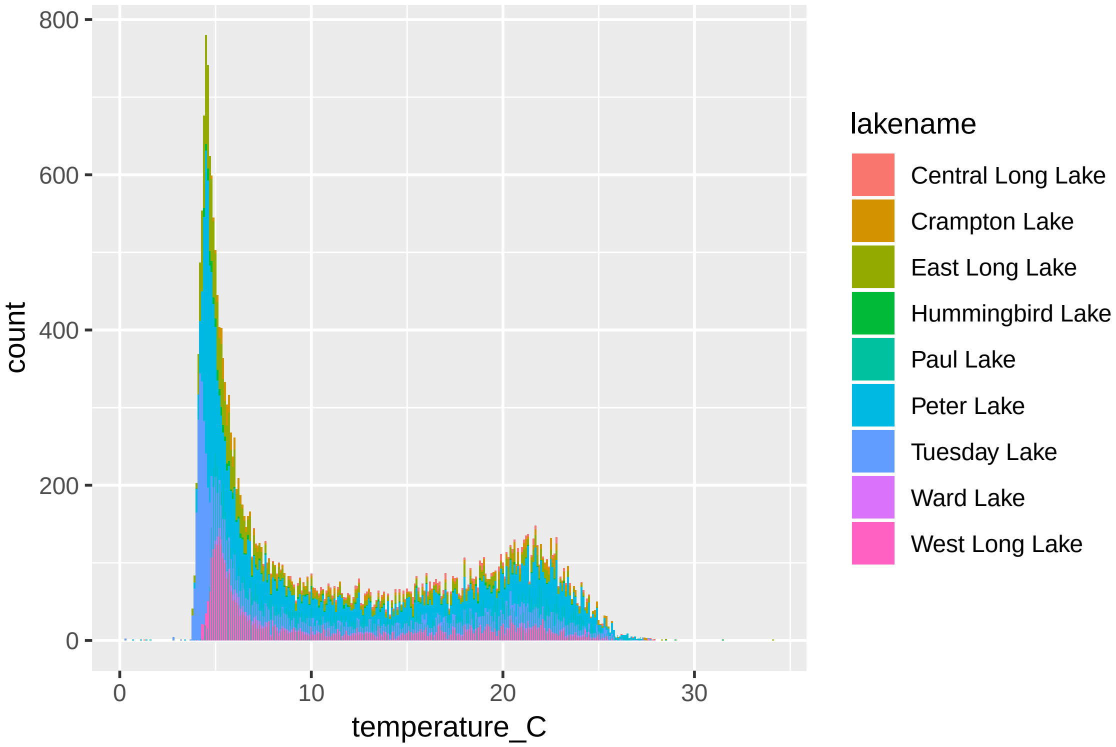

Figure 1: Bar chart of temperature for each lake

```
# 2 Histogram of count distributions of temperature
ggplot(lake) +
  geom_histogram(aes(x = temperature_C))
```

```
# 3 Change the bin
ggplot(lake) +
  geom_histogram(aes(x = temperature_C), binwidth = 0.5)
```

```
# 4 Frequency polygon of temperature for each lake
ggplot(lake) +
  geom_freqpoly(aes(x = temperature_C, color= lakename), bins = 50) +
  scale_color_brewer(palette="Set1")
```

```
# 5 Boxplot of temperature for each lake
ggplot(lake) +
  geom_boxplot(aes(x = lakename, y = temperature_C)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
# 6 Boxplot of temperature based on depth
ggplot(lake) +
  geom_boxplot(aes(x = depth, y = temperature_C, group = cut_width(depth, 0.25)))
```

```
# 7 Scatterplot of temperature by depth
ggplot(lake) +
  geom_point(aes(x = depth, y = temperature_C))
```

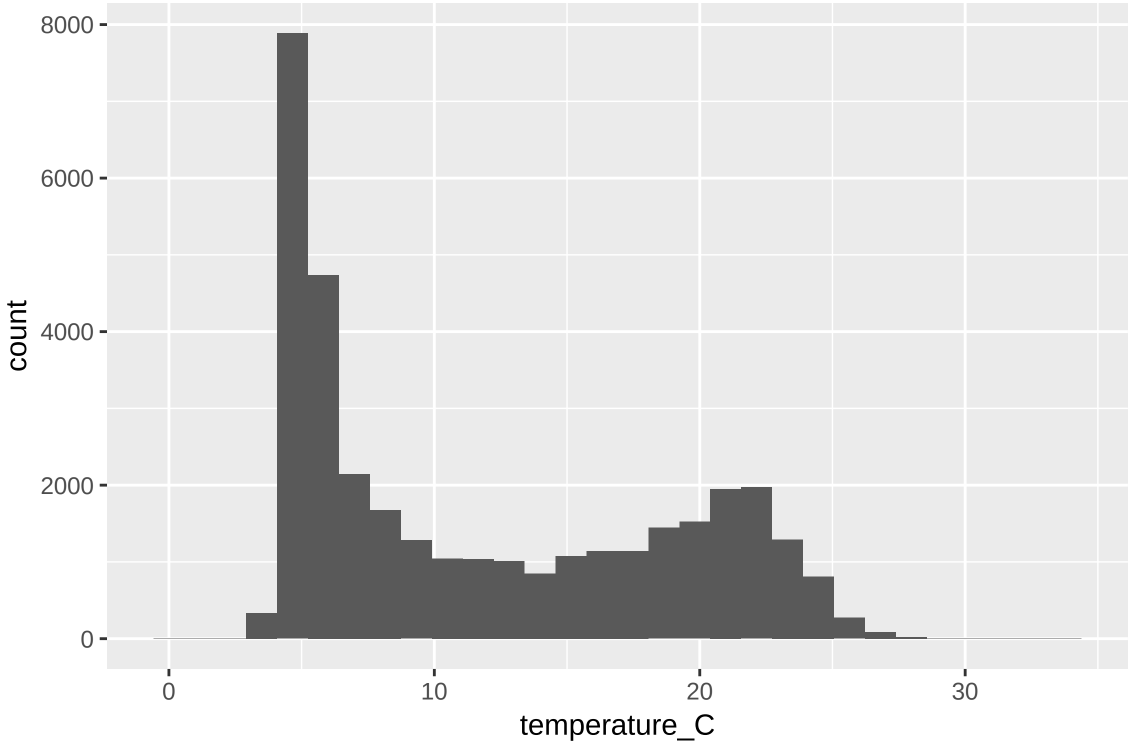

Figure 2: Histogram of distributions of temperature

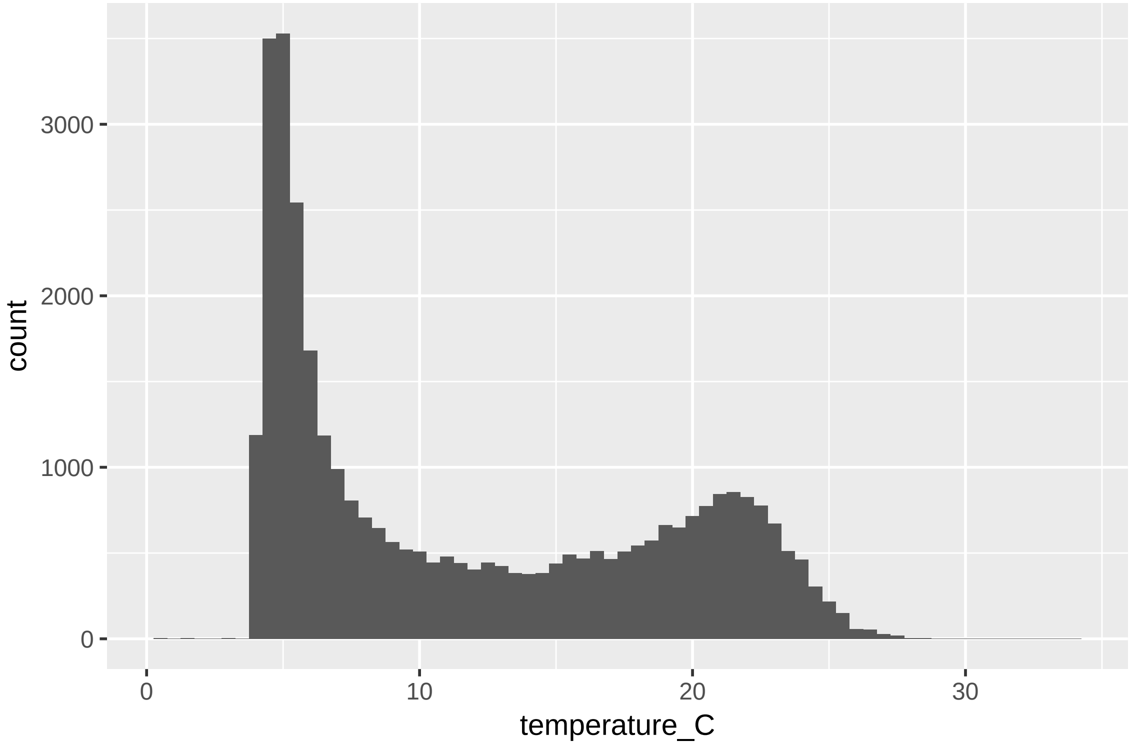

Figure 3: Histogram of detailed distributions of temperature

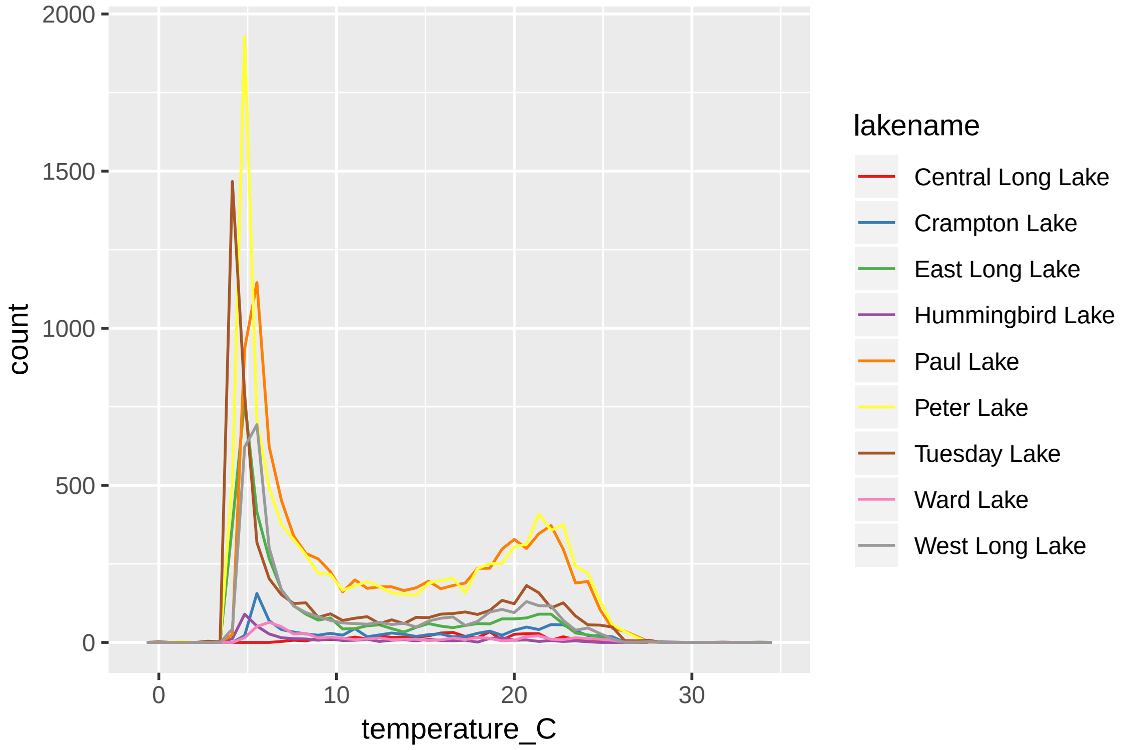

Figure 4: Frequency polygon of temperature for each lake

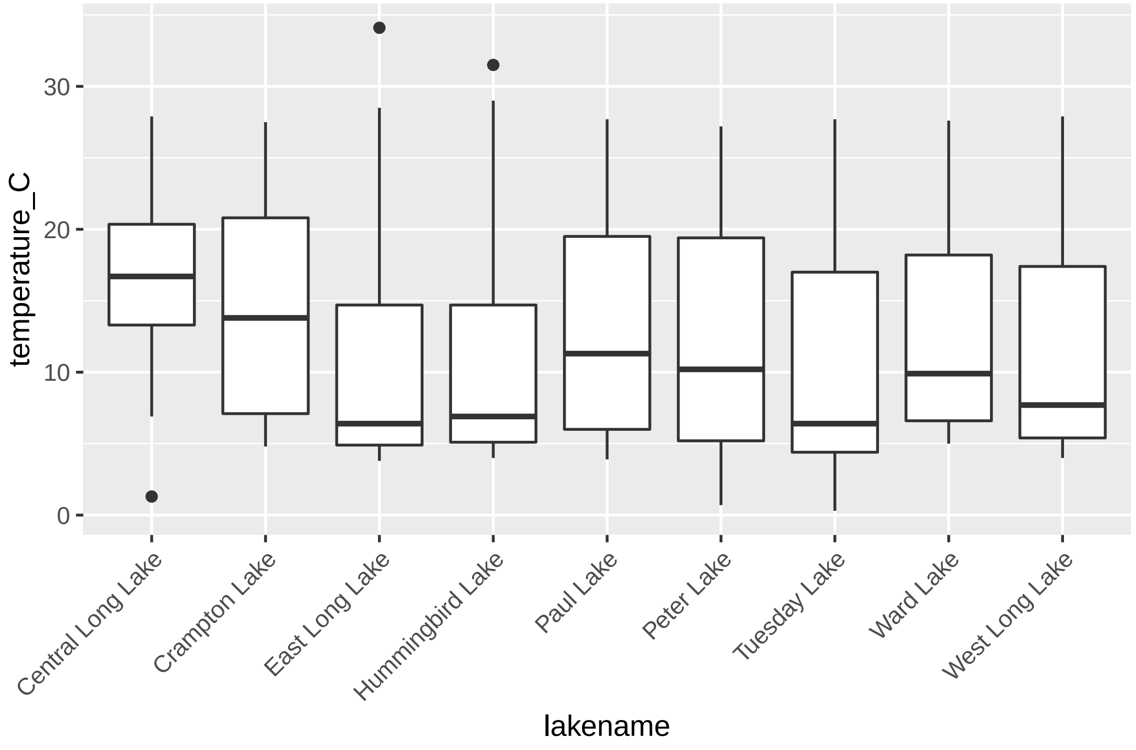

Figure 5: Boxplot of temperature for each lake

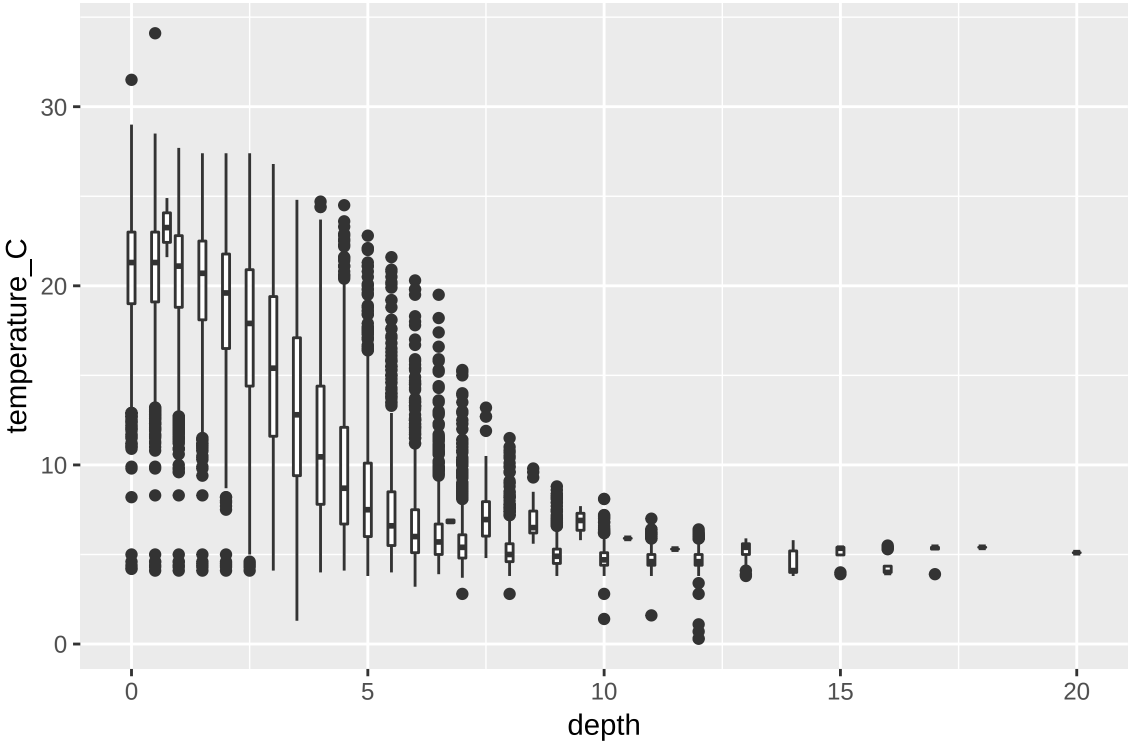

Figure 6: Boxplot of temperature based on depth

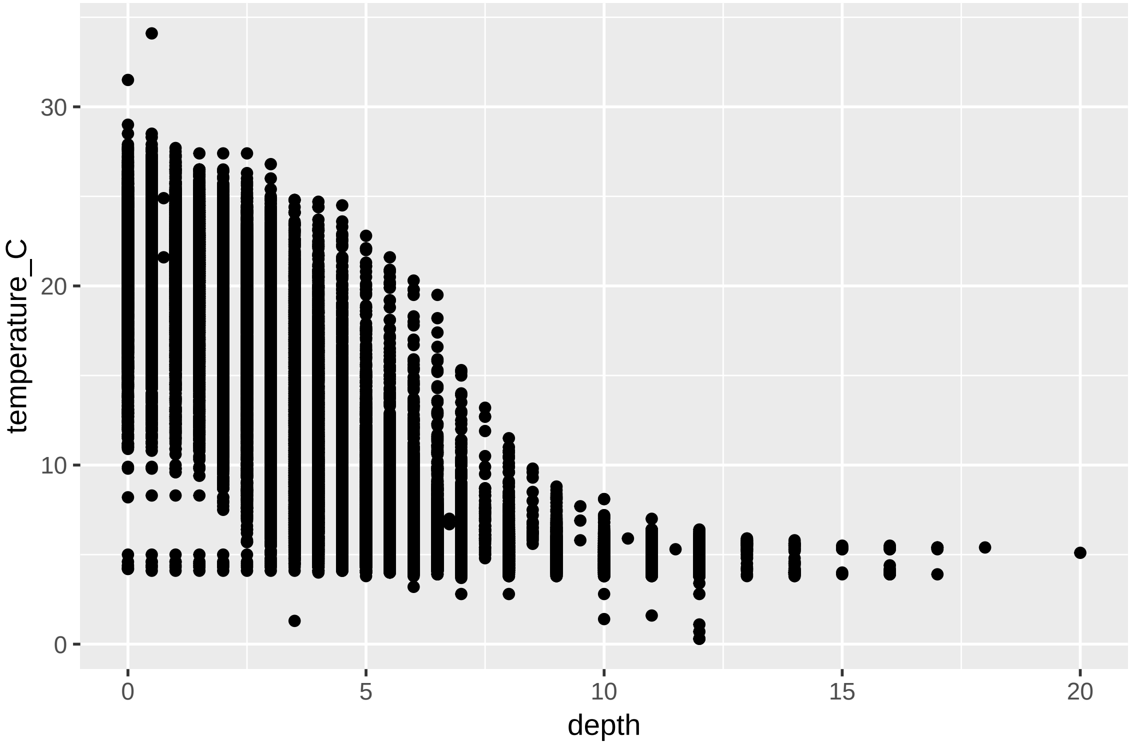

Figure 7: Scatterplot of temperature by depth

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: This dataset has 38614 observations and 11 variables, containing information such as temperature, depth, dissolved oxygen, in nine lakes in the North Temperate Lakes District in Wisconsin, USA. The data was collected from 1984 to 2016. From the graphs above, we may tell that though there are differences in the temperature between the nine lakes, they share the same trend that most of the counts fall in the range of 3 to 6 degree. East Long Lake has the overall lowest water temperature. Besides, there is a negative relationship between water temperature and depth.

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

> ANSWER 1: What are the dissolved oxygen concentrations in each lake?
>
> ANSWER 2: Is there any relationship between water depth and dissolved oxygen in each lake and in all lakes?
>
> ANSWER 3: Is there any relationship between water temperature and dissolved oxygen in each lake and in all lakes?