Assignment 3: Data Exploration

Caroline Reents

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()
```

```
## [1] "/Users/carolinereents/Desktop/Data Analytics/EnvironmentalDataAnalytics/Assignments"
```

```
setwd("~/Desktop/Data Analytics/EnvironmentalDataAnalytics")
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   0.7.8
## v tidyr   0.8.2     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()
```

```
library(readr)

lake_dat <- read_csv("./Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv")
```

```
## Parsed with column specification:
## cols(
```

```
##   lakeid = col_character(),
##   lakename = col_character(),
##   year4 = col_integer(),
##   daynum = col_integer(),
##   sampledate = col_character(),
##   depth = col_double(),
##   temperature_C = col_double(),
##   dissolvedOxygen = col_double(),
##   irradianceWater = col_double(),
##   irradianceDeck = col_integer(),
##   comments = col_character()
## )
```

```
## Warning in rbind(names(probs), probs_f): number of columns of result is not
## a multiple of vector length (arg 1)
```

```
## Warning: 3232 parsing failures.
## row # A tibble: 5 x 5 col     row col         expected        actual file
## ... ................. ... ..........................................................................
## See problems(...) for more details.
```

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: I learned that the data was collected in Wisconsin. I also learned how each of the variables were collected and what years the different pieced were collected during. (Carbon: 84-16, nutrients 91-16, limnology 84-16).

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(lake_dat)
```

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

```
# 2
class(lake_dat)
```

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

```
# 3
head(lake_dat, 8)
```

```
## # A tibble: 8 x 11
##   lakeid lakename year4 daynum sampledate depth temperature_C
##   <chr>  <chr>    <int>  <int> <chr>      <dbl>         <dbl>
## 1 L      Paul La~  1984    148 5/27/84     0             14.5
```

```
## 2 L      Paul La~  1984    148 5/27/84     0.25          NA
## 3 L      Paul La~  1984    148 5/27/84     0.5           NA
## 4 L      Paul La~  1984    148 5/27/84     0.75          NA
## 5 L      Paul La~  1984    148 5/27/84     1             14.5
## 6 L      Paul La~  1984    148 5/27/84     1.5           NA
## 7 L      Paul La~  1984    148 5/27/84     2             14.2
## 8 L      Paul La~  1984    148 5/27/84     3             11
## # ... with 4 more variables: dissolvedOxygen <dbl>, irradianceWater <dbl>,
## #   irradianceDeck <int>, comments <chr>
```

```
# 4
class(lake_dat$lakename)
```

```
## [1] "character"
```

```
class(lake_dat$sampledate)
```

```
## [1] "character"
```

```
class(lake_dat$depth)
```

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

```
class(lake_dat$temperature_C)
```

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

```
# 5
summary(lake_dat$lakename)
```

```
##    Length     Class      Mode
##     38614 character character
```

```
summary(lake_dat$depth)
```

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

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

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

Question: Do you want to remove NAs from this dataset? Why or why not?

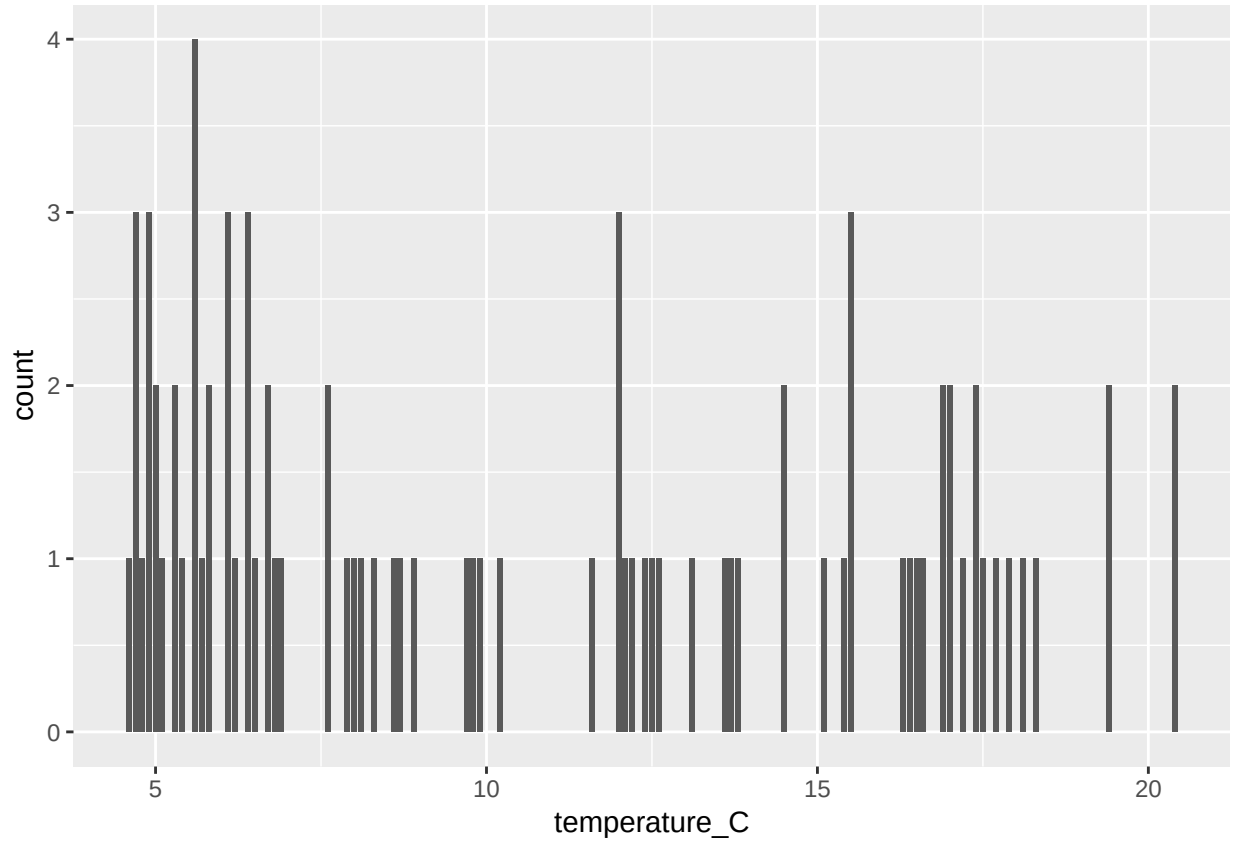
> ANSWER: I would remove the NAs from the data set because I am doing quantitative analysis comparing variables, so it is necessary that each measure have all of the variables otherwise certain analyses will not work.
>
> note, when you remove the NAs a lot of the lakes disappear from the data set, so depending on your research goals it may or may not be necessarry to remove the NAs. Looking at the data, most NAs are not in the depth or temp variables, and since these are the data we are looking at, I would not suggest removing NAs for this particular assignment. But in general, I would remove thee NAs, so I kept my analysis to a data set without NAs.

4) Explore your data graphically

Write R commands to display graphs depicting:

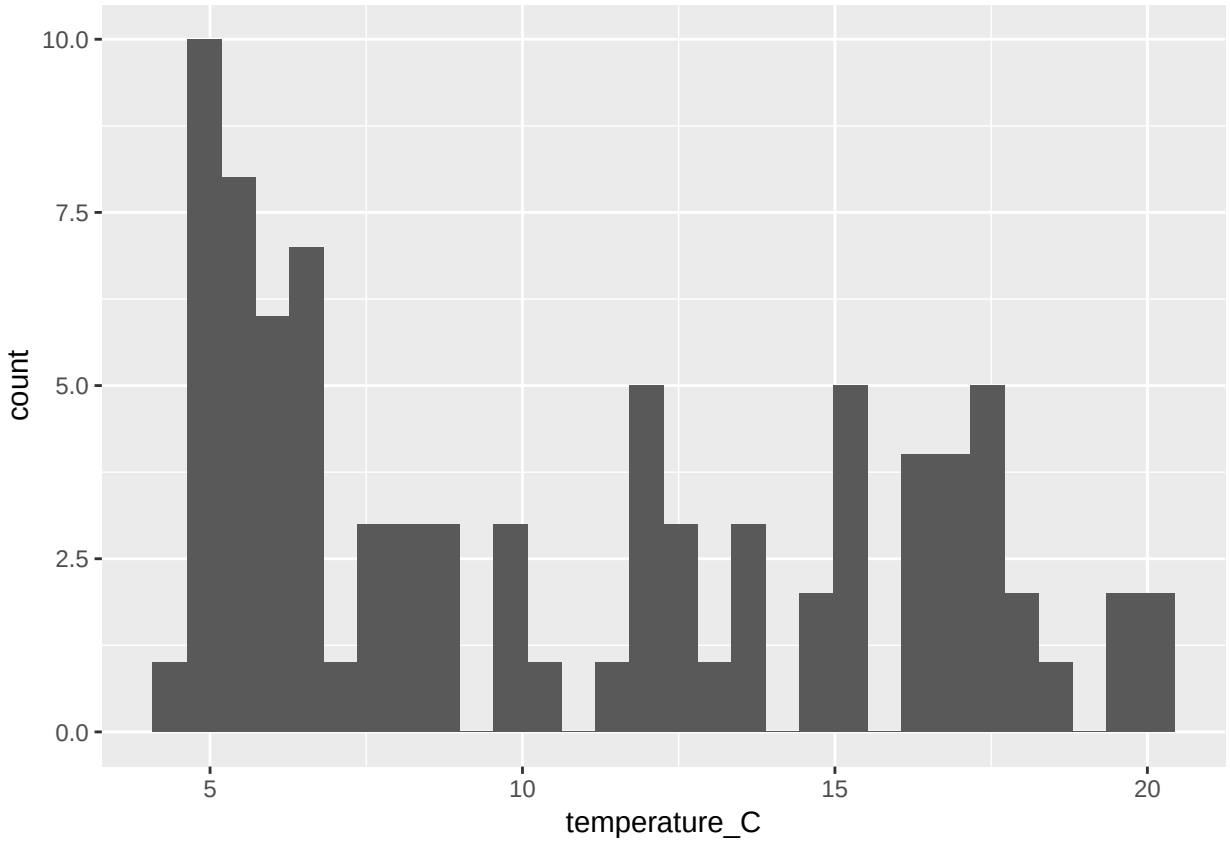
1. Bar chart of temperature counts for each lake
2. Histogram of count distributions of temperature (all temp measurements together)
3. Change histogram from 2 to have a different number or width of bins
4. Frequency polygon of temperature for each lake. Choose different colors for each lake.
5. Boxplot 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

```
lake_dat_full<-na.omit(lake_dat)
# 1
ggplot(lake_dat_full, aes(x = temperature_C)) +
  geom_bar()
```

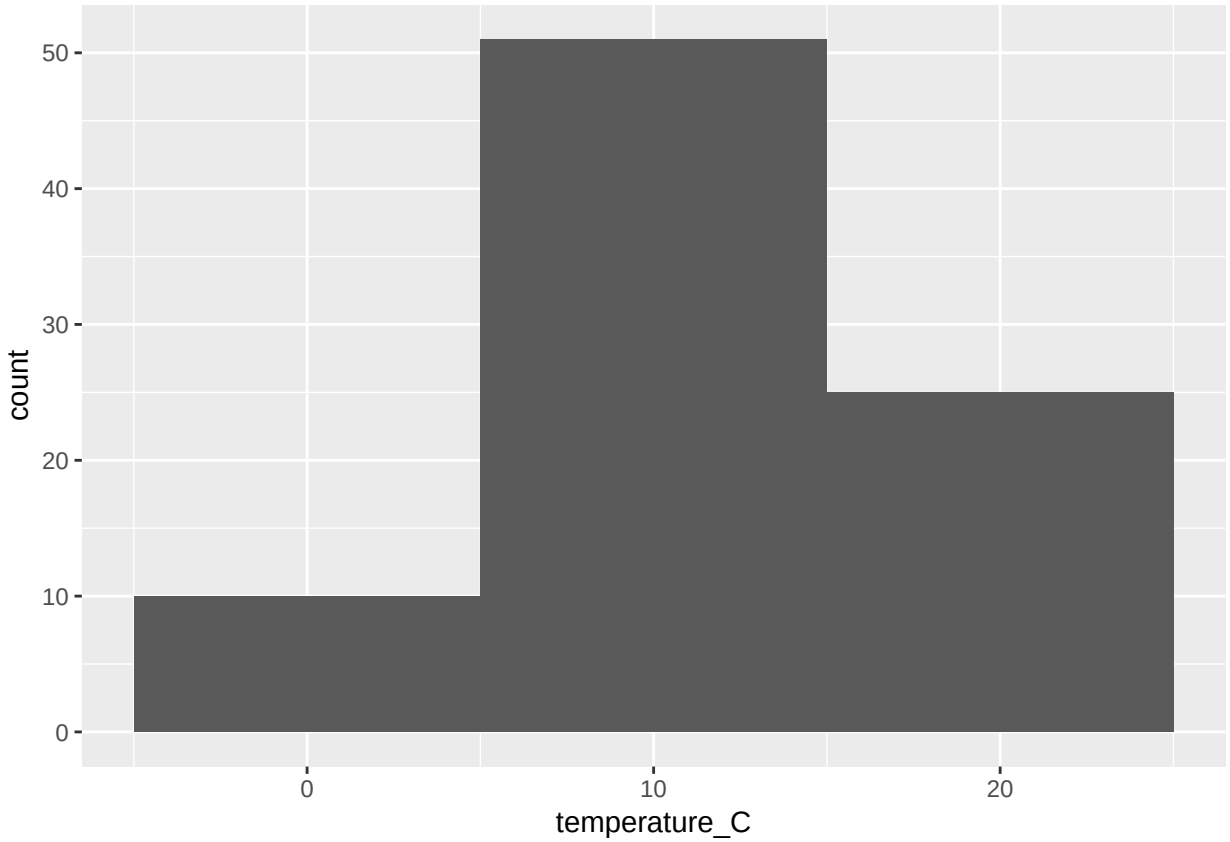



```
# 2
ggplot(lake_dat_full) +
  geom_histogram(aes(x = temperature_C))
```

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

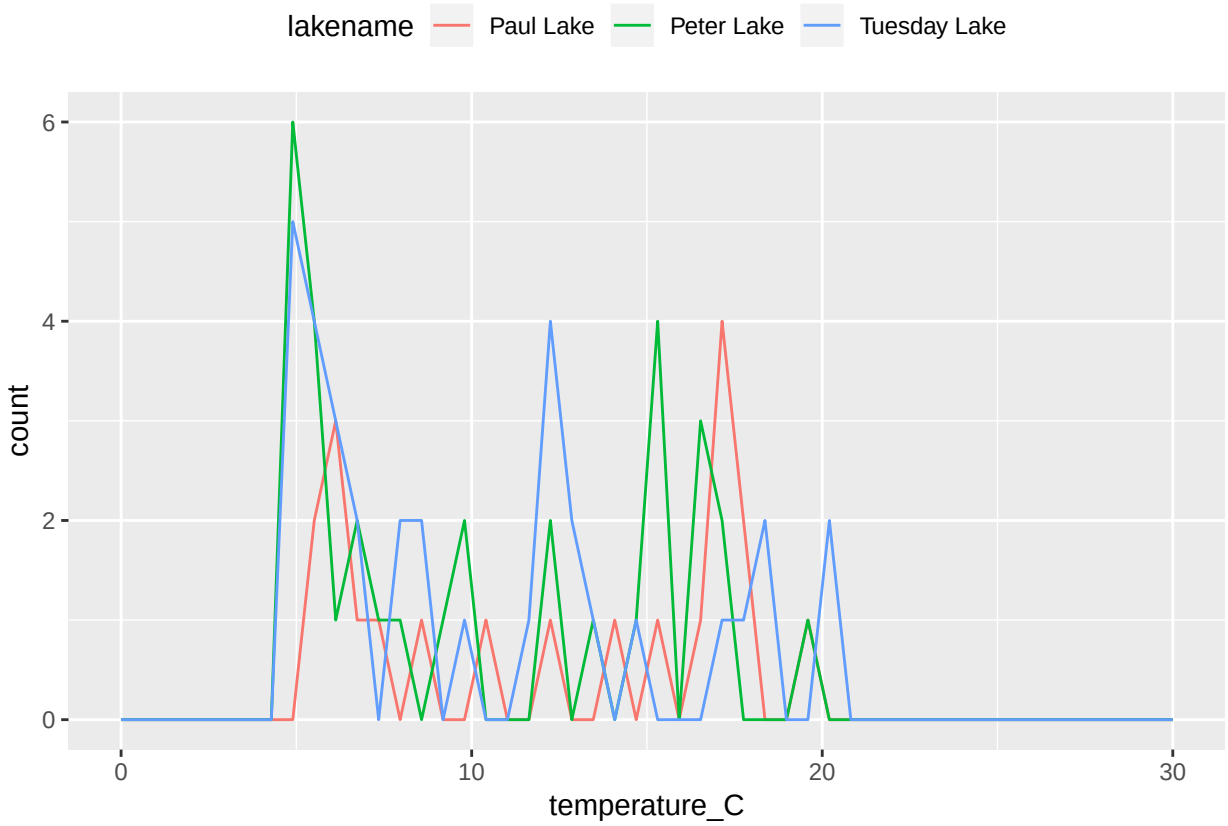



```
# 3
ggplot(lake_dat_full) +
  geom_histogram(aes(x = temperature_C), binwidth = 10)
```

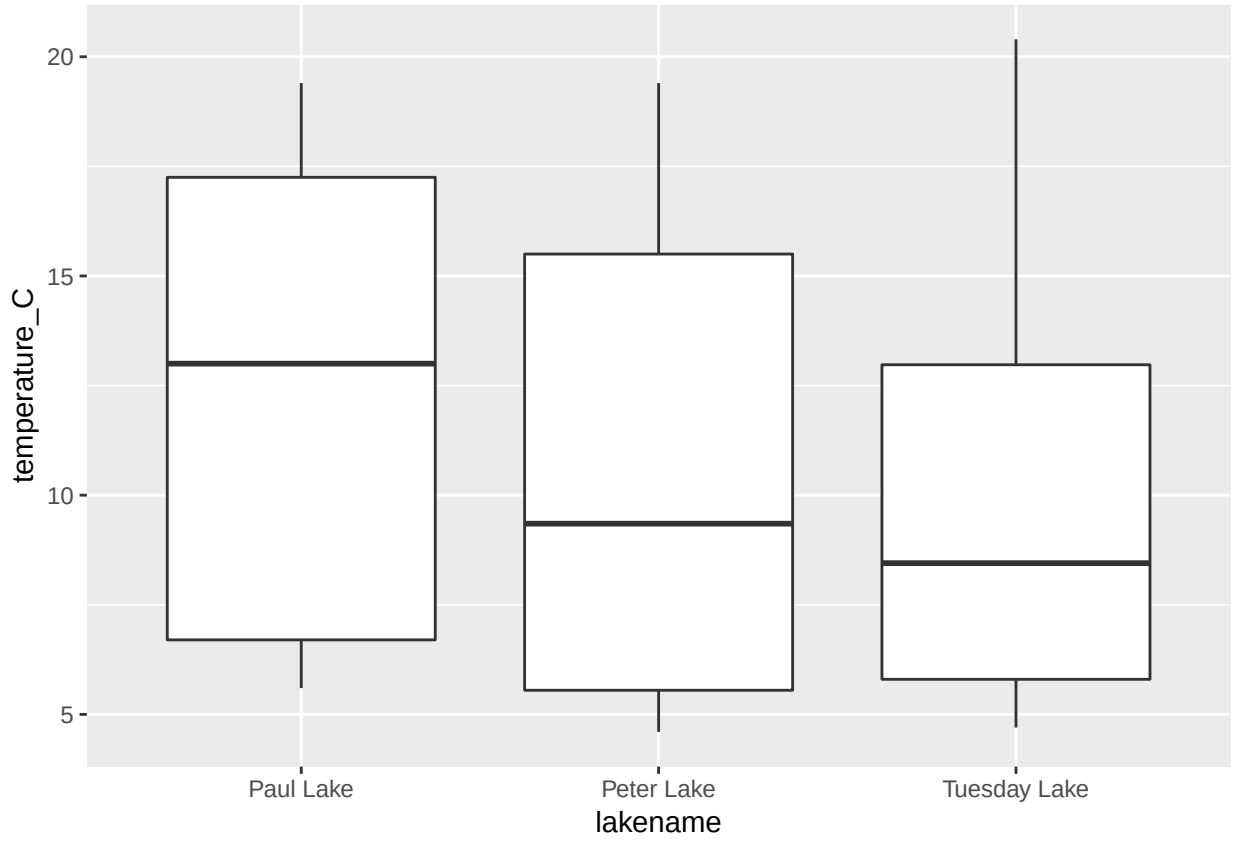



```
# 4
ggplot(lake_dat_full) +
  geom_freqpoly(aes(x = temperature_C, color = lakename), bins = 50) +
  scale_x_continuous(limits = c(0, 30)) +
  theme(legend.position = "top")
```

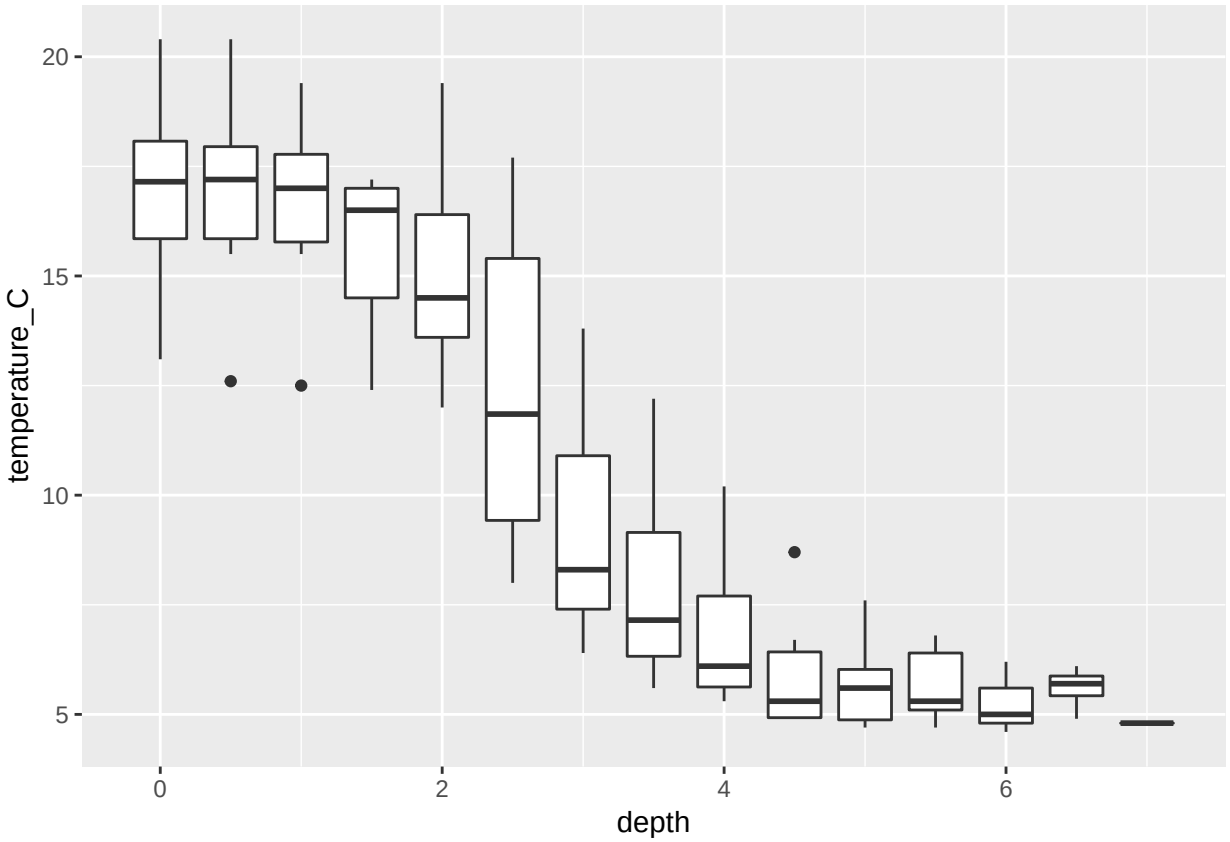
```
## Warning: Removed 6 rows containing missing values (geom_path).
```



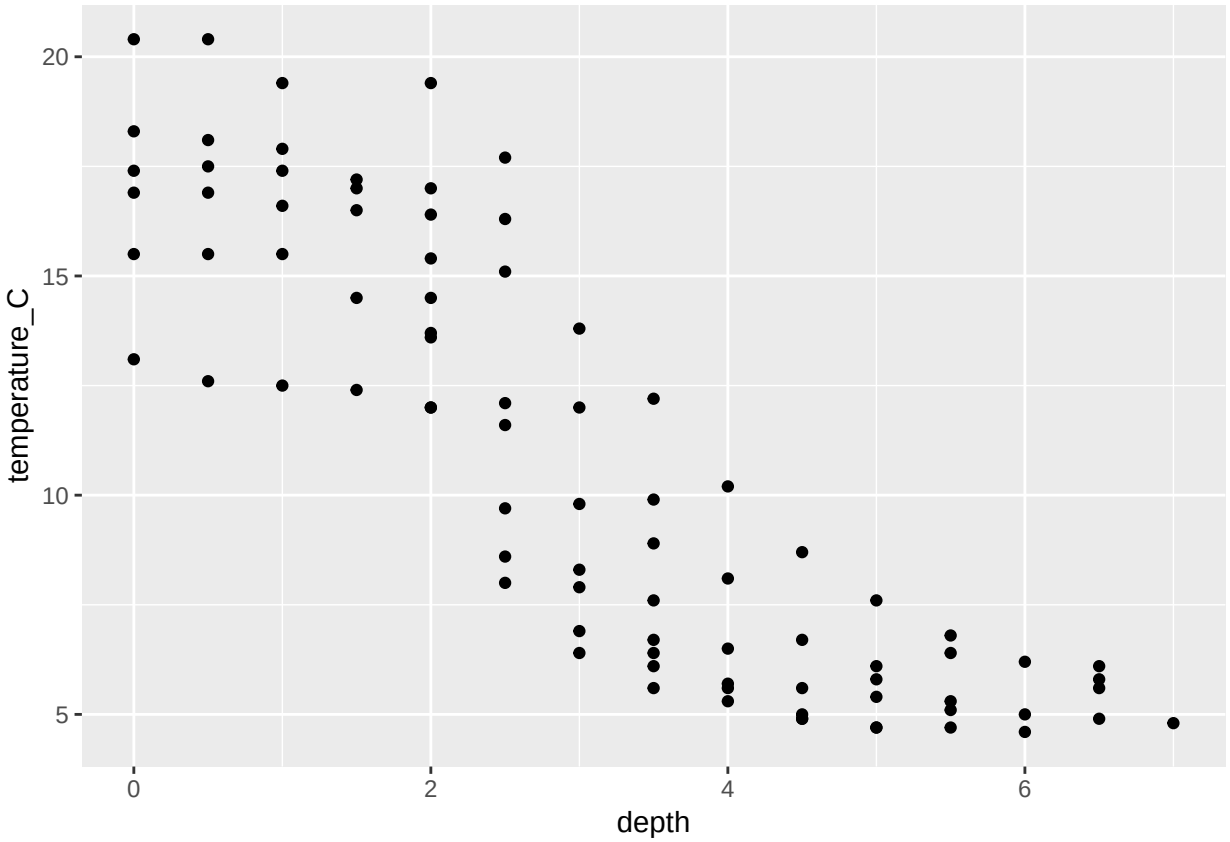

```
# 5
ggplot(lake_dat_full) +
  geom_boxplot(aes(x = lakename, y = temperature_C))
```




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




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



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: I found that most of the lake temperatures measured were around 5 degrees C. Also, for the most part, the lakes followed similar patterns in the number of counts for each temperature. Paul Lake has the highest average temperature of all of the lakes, but this difference is not significant. As to be expected, the graphs show that as depth increases, temperature decreases.

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

> ANSWER 1: I would be interested to see the temperature counts based on year, to see if there are large annual changes in temp.
>
> ANSWER 2: Depth based on lake name, it may be useful to know which lakes are deeper than others.
>
> ANSWER 3: How does dissolved oxygen in the lakes change based on temperature.