Assignment 3: Data Exploration

Sarah Ko

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 the working directory
getwd()
```

```
## [1] "C:/Users/Sarah/Documents/Duke/Year 2/Spring 2019/Data Analytics/Environmental_Data_Analytics/As
```

```
# Load necessary package 'tidyverse'
library(tidyverse)
```

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

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

```
## 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
```

```
## 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 'dplyr' was built under R version 3.5.2
```

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

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

```
# set wd to the filepath of Environmental_Data_Analytics/Assignments to use relative filepath
# upload the North Temperate Lakes long term monitoring dataset using relative filepath
NTL.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. The data contains information on Physical and Chemical Limnology, from the timespan of 1984 - 2016
2. This data was collected from 1 central station near the deepest point of each lake
3. When chemical measurements were made in vertical profiles, depths for sampling usually correspond to the surface plus depths of 50%, 25%, 10%, 5% and 1% of the surface irradiance

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.data)
```

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

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

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

```
# 3
head(NTL.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(NTL.data$lakename)
```

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

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

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

```
class(NTL.data$depth)
```

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

```
class(NTL.data$temperature_C)
```

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

```
# 5
summary(NTL.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
```

```
summary(NTL.data$depth)
```

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

```
summary(NTL.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.

```
# change class of sampledate from factor to date. must define original format of data
NTL.data$sampledate <- as.Date(NTL.data$sampledate, format = "%m/%d/%y")

# confirm that the class is date
class(NTL.data$sampledate)
```

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

```
#display the first 10 rows of the date column
head(NTL.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: Removing or keeping NAs depends on the research question. If I suspected that the data is unavailable due to another variable, it would be good to keep NAs in the dataset - deleting them may create bias. If I suspected that this data was unavailable at random and I had adequate

amounts of data to create my model, I may be inclined to delete the NAs. Another reason to delete the NAs is if the data was unavailable for a predictable reason (e.g. incomplete years like the beginning and ending years of a dataset may have missing data).

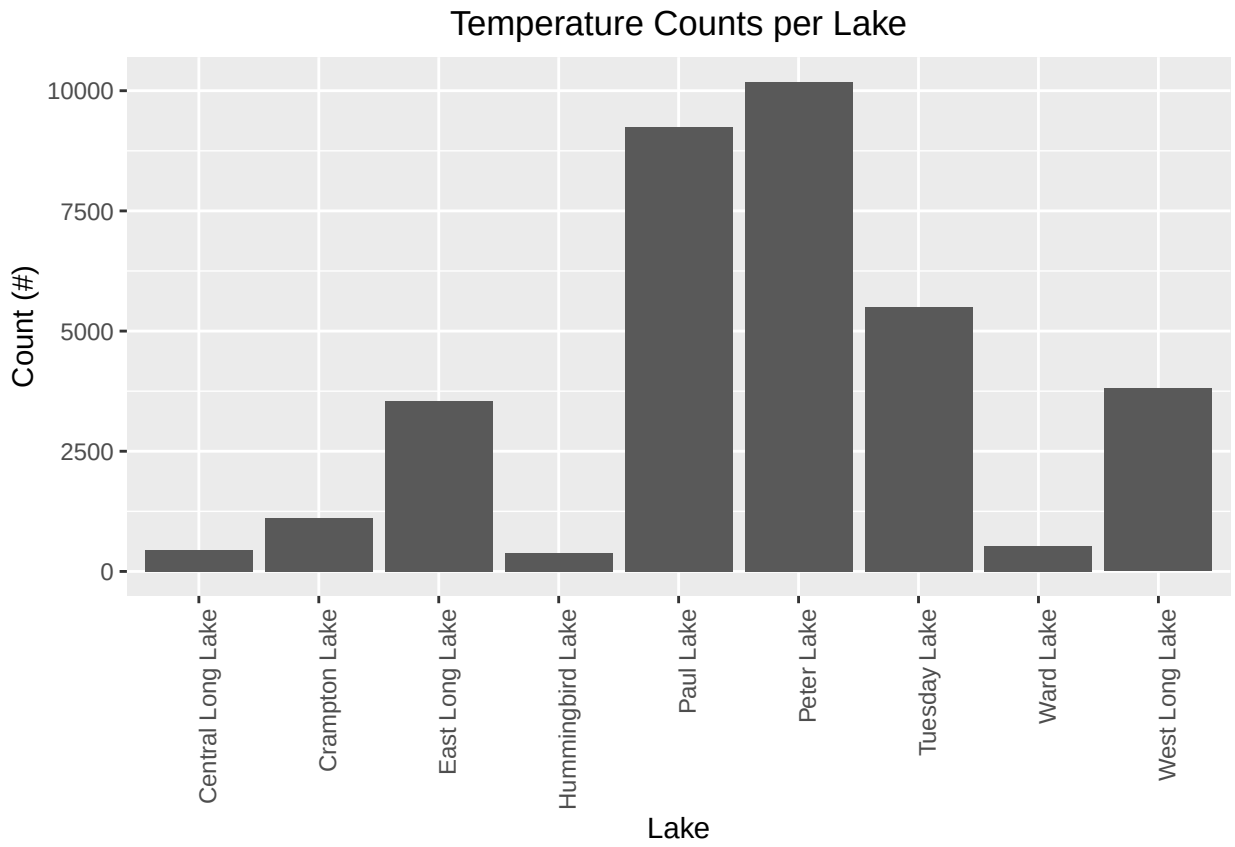
For this particular dataset, I would want to remove the NAs in order to perform statistical analyses.

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

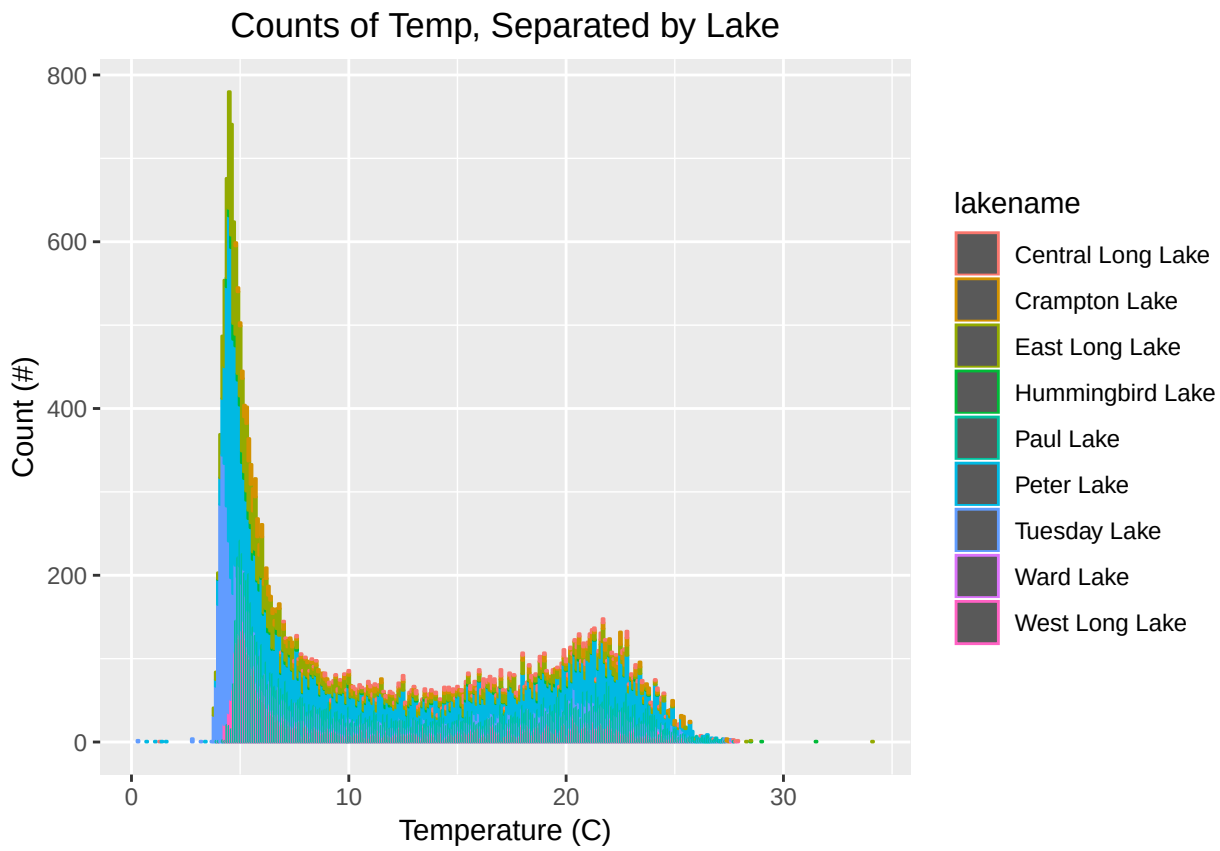
```
# 1

# create dataframe with only lakename and temperature data
NTL.lakename.temp <- NTL.data[,c("lakename", "temperature_C")]

# remove NAs
NTL.lakename.temp.complete <- na.omit(NTL.lakename.temp)

# barplot counts the occurance of each lake, which each has a corresponding temperature record
ggplot(data = NTL.lakename.temp.complete, aes(x = lakename)) + geom_bar() +
  ggtitle("Temperature Counts per Lake") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "Lake", y = "Count (#)") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




```
# create second barplot of temperature with different colors for lakename
ggplot(data = NTL.data, aes(x = temperature_C, color=lakename)) + geom_bar() +
  ggtitle("Counts of Temp, Separated by Lake") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "Temperature (C)", y = "Count (#)")
```

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

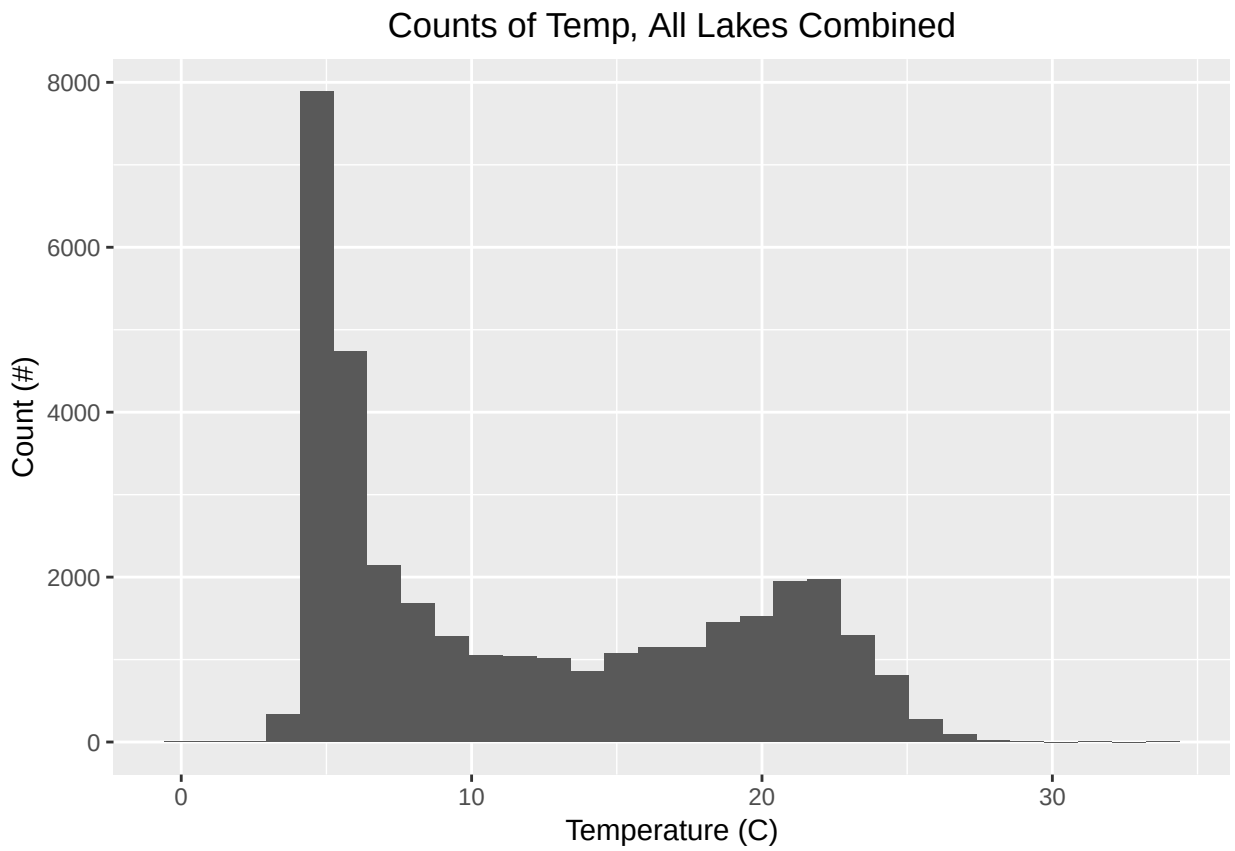
```
## Warning: position_stack requires non-overlapping x intervals
```




```
# 2

# histogram of temp
ggplot(NTL.data) + geom_histogram(aes(x = temperature_C)) +
  ggtitle("Counts of Temp, All Lakes Combined") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "Temperature (C)", y = "Count (#)")
```

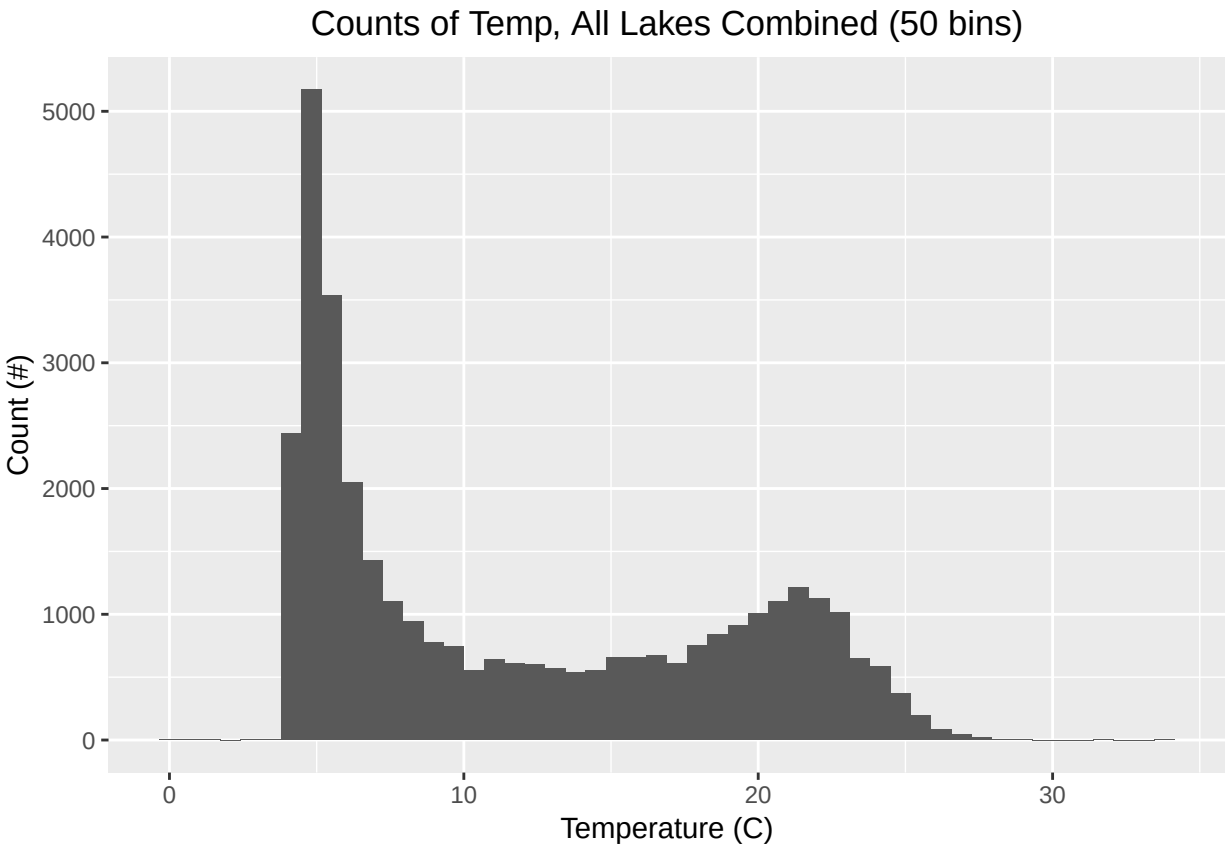
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

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




```
# 3

# create histogram with 50 bins
ggplot(NTL.data) +
  geom_histogram(aes(x = temperature_C), bins=50)+
  ggtitle("Counts of Temp, All Lakes Combined (50 bins)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "Temperature (C)", y = "Count (#)")
```

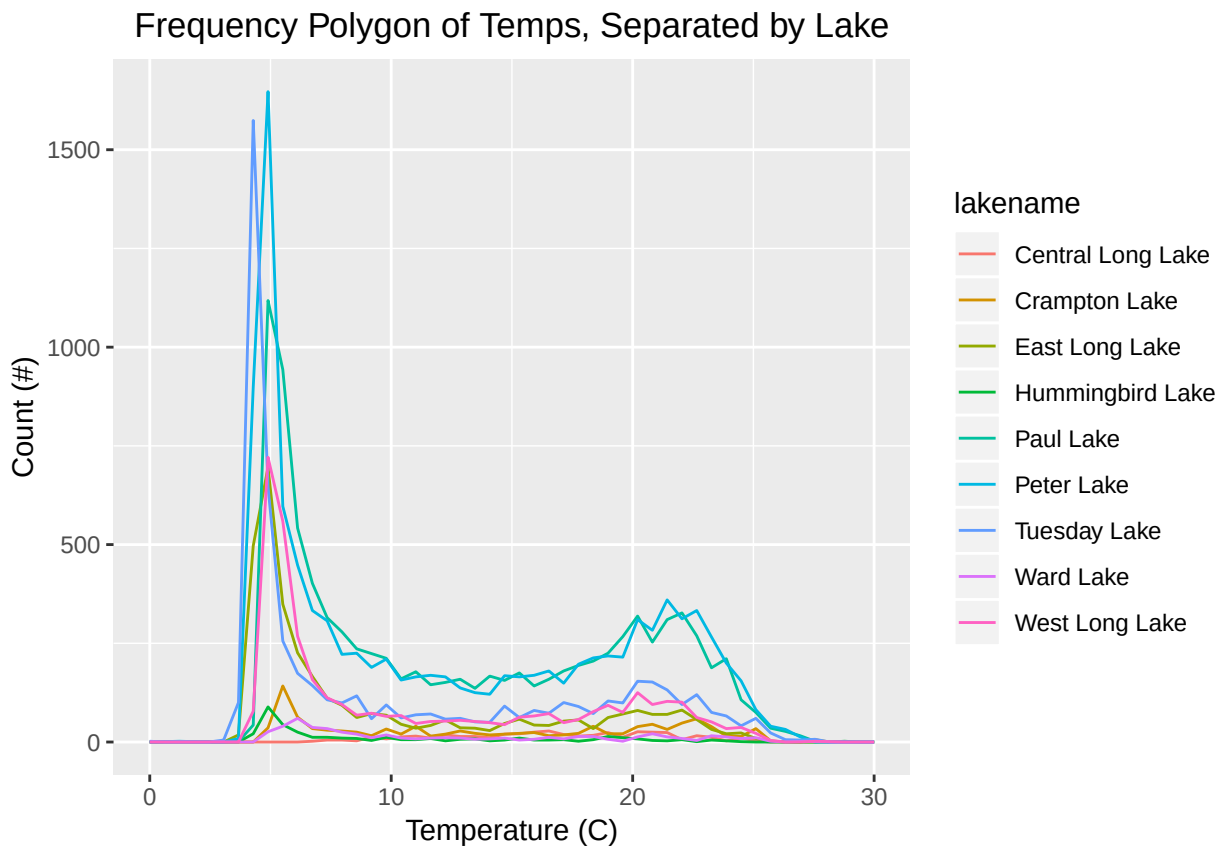
```
## Warning: Removed 3858 rows containing non-finite values (stat_bin).
```




```
# 4

# frequency polygon of temp
ggplot(NTL.data) +
  geom_freqpoly(aes(x = temperature_C, color = lakename), bins = 50) +
  scale_x_continuous(limits = c(0, 30)) +
  theme(legend.position = "right") +
  ggtitle("Frequency Polygon of Temps, Separated by Lake") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "Temperature (C)", y = "Count (#)")
```

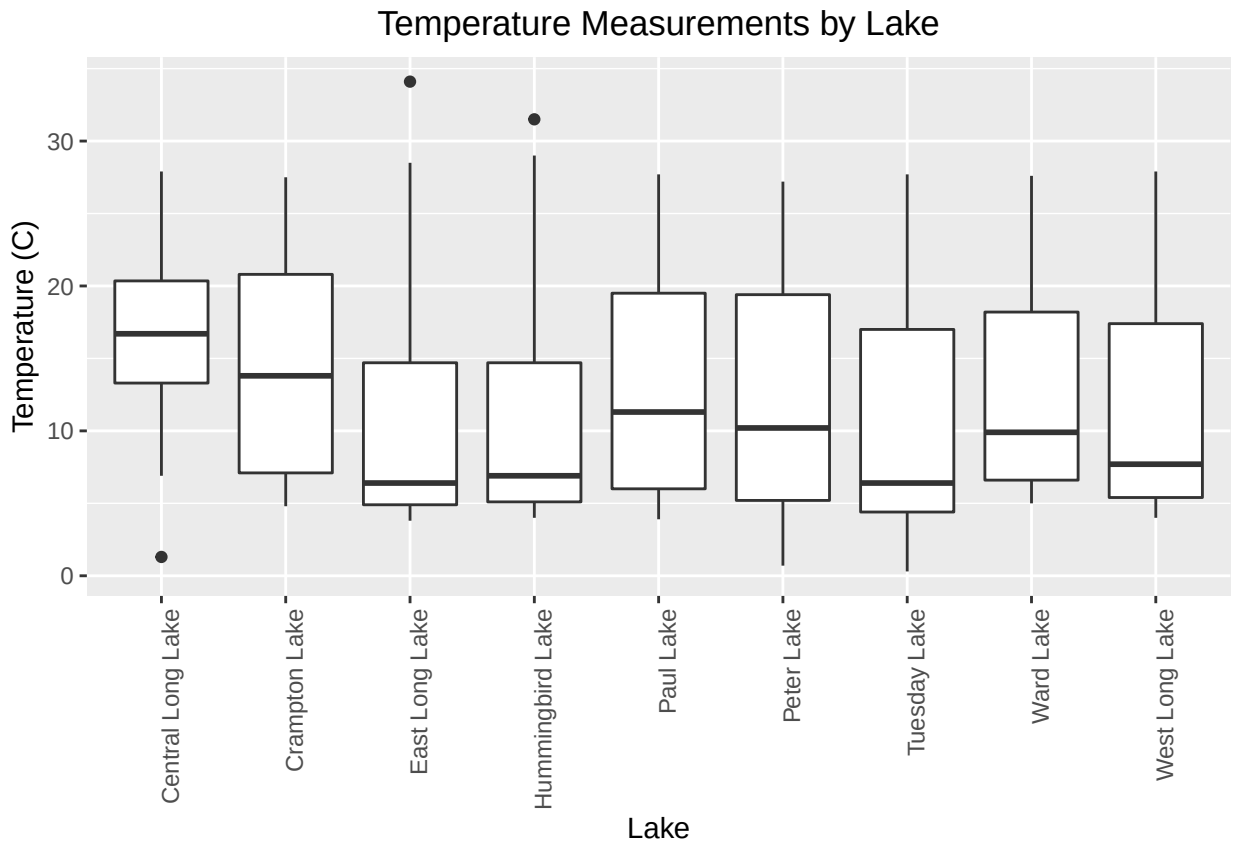
```
## Warning: Removed 3860 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 18 rows containing missing values (geom_path).
```



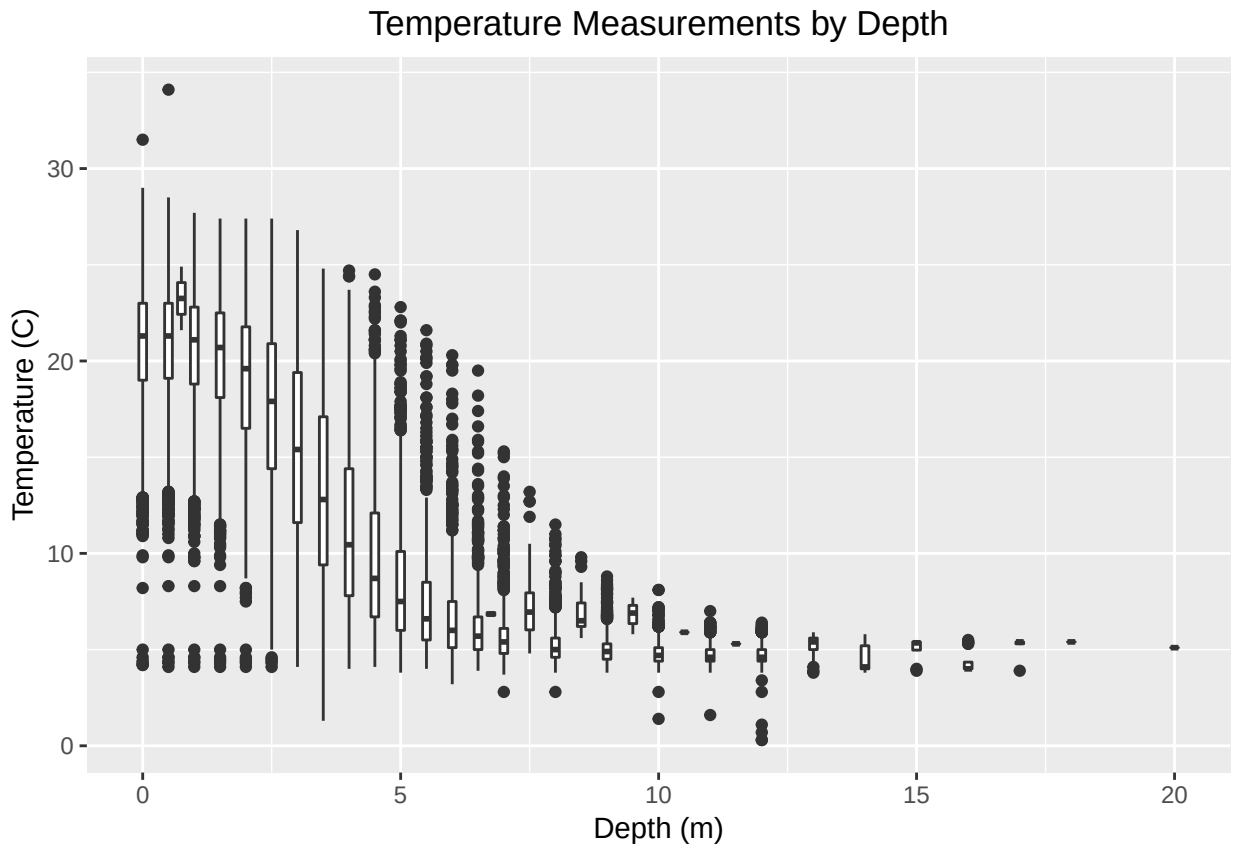

```
# 5

# boxplot of temp per lake
ggplot(NTL.data) +
  geom_boxplot(aes(x = lakename, y = temperature_C, group = cut_width(lakename, 1))) +
ggtitle("Temperature Measurements by Lake") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "Lake", y = "Temperature (C)") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

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



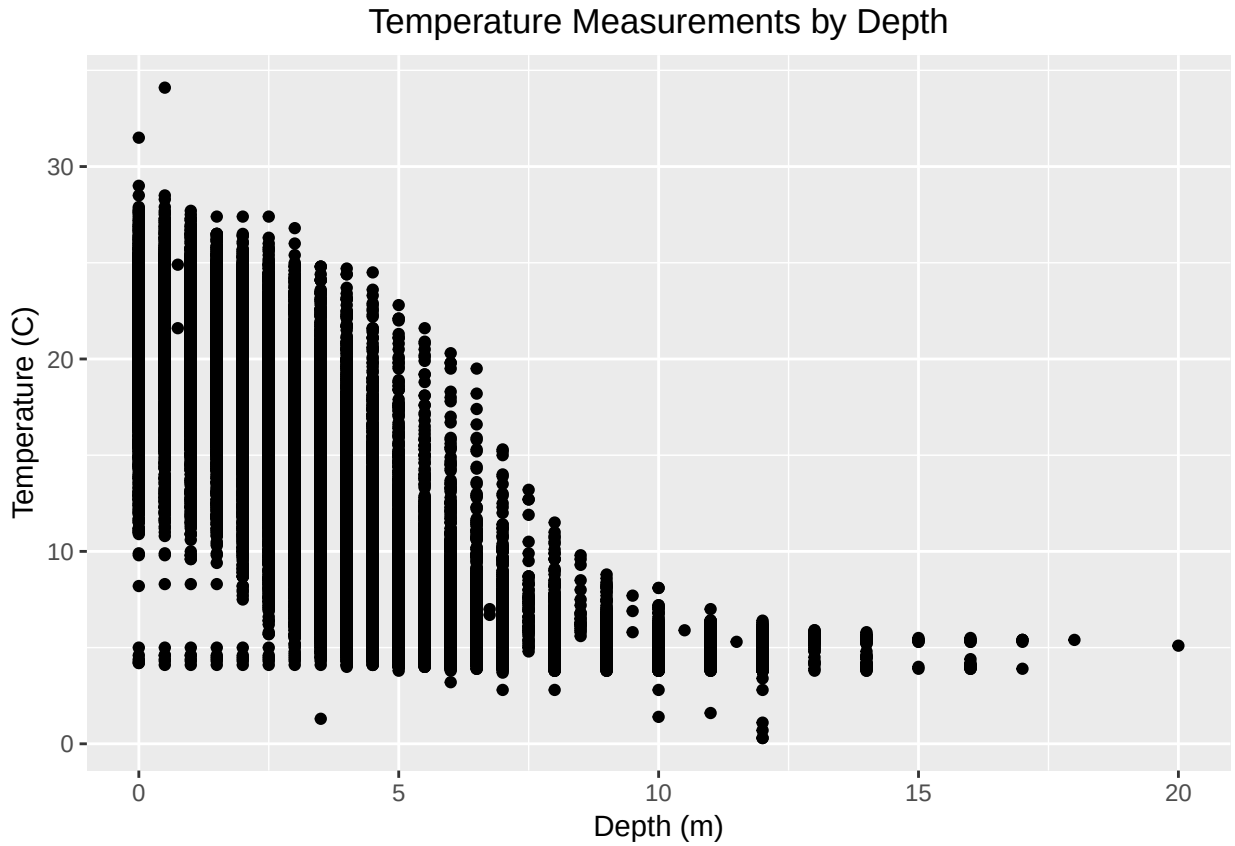

```
# 6

# boxplot of temp based on depth, depth divided into 0.25m increments
ggplot(NTL.data) +
  geom_boxplot(aes(x = depth, y = temperature_C, group = cut_width(depth, 0.25))) +
  ggtitle("Temperature Measurements by Depth") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "Depth (m)", y = "Temperature (C)")
```

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




```
# 7

# scatterplot temp based on depth
ggplot(NTL.data) +
  geom_point(aes(x = depth, y = temperature_C)) +
  ggtitle("Temperature Measurements by Depth") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "Depth (m)", 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: The bar plot of temperature readings showed that Paul Lake and Peter Lake have by far the most temperature measurements, followed by Tuesday Lake, West Long Lake, and East Long Lake. Temperature measurements compiled from all the lakes show a range of 3-28 C. Most of the temperature measurements are around 5 C, with a small cluster of measurements around 23 C. At shallower depths (~0-7m), the temperature measurements have a much larger distribution as compared to those deeper. Even so, the temperature appears to decrease with increasing depth below the surface.

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

> ANSWER 1: What is the mathematical correlation between temperature and depth?
>
> ANSWER 2: Are the temperature means of the lakes significantly different from one another?
>
> ANSWER 3: Is there a correlation between temperature and number of temperature measurements recorded?