Assignment 3: Data Exploration

Sebastian Bognar

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
# check working directory

getwd()
```

```
## [1] "/Users/Seabass/Documents/Duke/spring_2019/env_872L/lesson_2/ENV_872L/Assignments"
```

```r
# load tidyverse package

library(tidyverse)
```

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

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

```r
# upload csv for north temperate lakes

temp_L_raw<-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 first salient piece of information from the readme file that I learned was that the data was collected from several lakes in the North Temperate Lakes District in Wisconsin. I also learned that the data was collected from 1984-2016, which is important. I also learned that the physical and chemical variables were measured at central stations near the deepest part of each lake.

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 display the dimensions of the dataset
dim(temp_L_raw)
```

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

```
# 2 display the class of the dataset
class(temp_L_raw)
```

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

```
# 3 display the first 8 rows of the dataset

head(temp_L_raw, 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 display the class of the variables:lakename, sampledate, depth, and temperature

class(temp_L_raw$lakename)
```

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

```r
class(temp_L_raw$sampledate)
```

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

```r
class(temp_L_raw$depth)
```

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

```r
class(temp_L_raw$temperature_C)
```

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

```r
# 5 display summaries of lakename, depth, and temperature

summary(temp_L_raw$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(temp_L_raw$depth)
```

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

```r
summary(temp_L_raw$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
# convert sampledate class from factor to date

temp_L_raw$sampledate <- as.Date(temp_L_raw$sampledate, format = "%m/%d/%y")

temp_L_raw$sampledate <- format(temp_L_raw$sampledate,format = "%m/%d/%y")

# show first 10 rows of data column

head(temp_L_raw$sampledate,10)
```

```
##  [1] "05/27/84" "05/27/84" "05/27/84" "05/27/84" "05/27/84" "05/27/84"
##  [7] "05/27/84" "05/27/84" "05/27/84" "05/27/84"
```

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

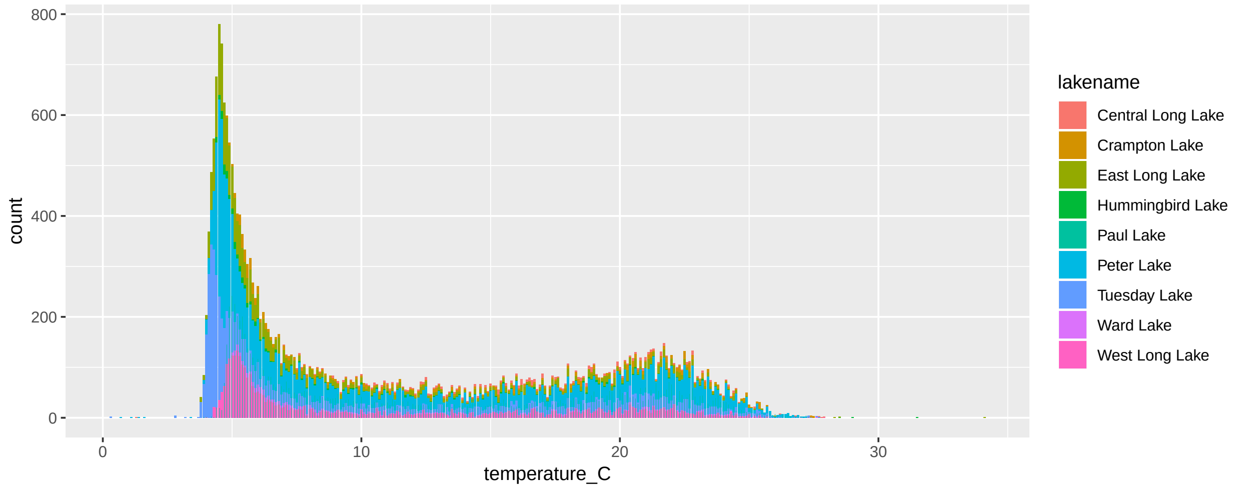
> ANSWER: For this assignment, the removal of NAs is not necessary because we are not doing any statistical analysis that would require the removal of the NAs. Additionally, we do not know what the NAs represent because it was not stated in the readme document. If they represented levels below detection it maybe warranted to leave them in, but if they are just no sampling or data then removing them would be fine.

4) Explore your data graphically

Write R commands to display graphs depicting:

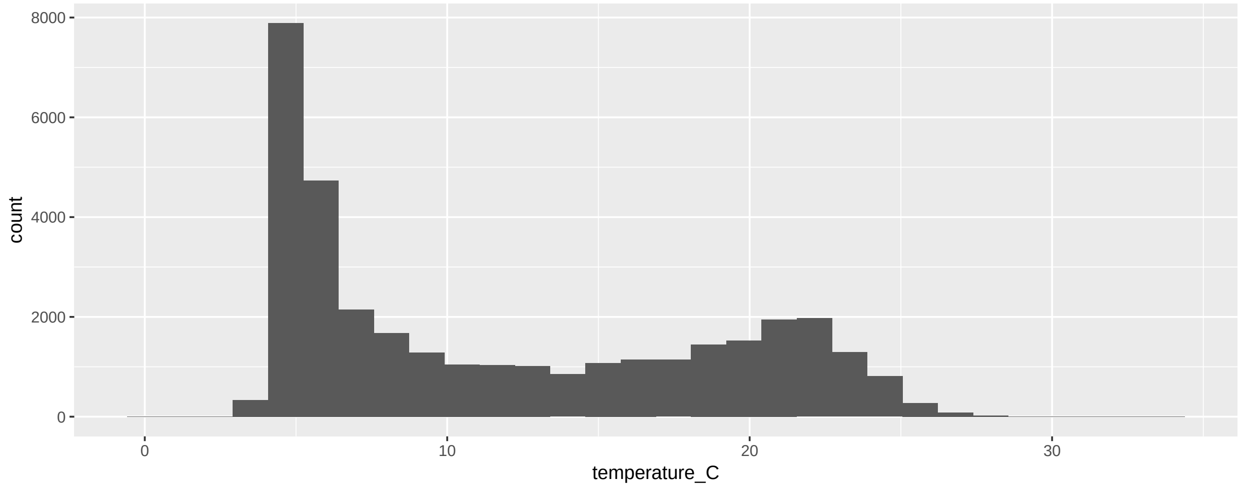
1. Bar chart of temperature counts for each lake
2. Histogram of count distributions of temperature (all temp measurements together)
3. Change histogram from 2 to have a different number or width of bins
4. Frequency polygon of temperature for each lake. Choose different colors for each lake.
5. Boxplot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

```
# load ggplot

library(ggplot2)
# 1 bar chart of temp counts for each lake

ggplot(temp_L_raw, aes(temperature_C)) +
  geom_bar(aes(fill = lakename))
```

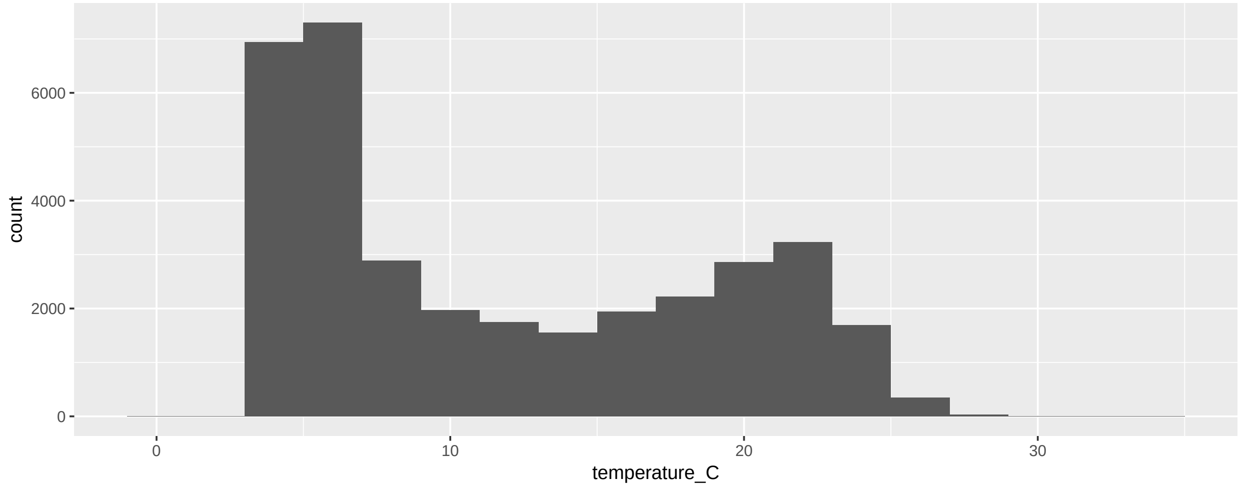


```
# 2 Histogram of count distributions of temperature

ggplot(temp_L_raw) +
  geom_histogram(aes(x = temperature_C))
```

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

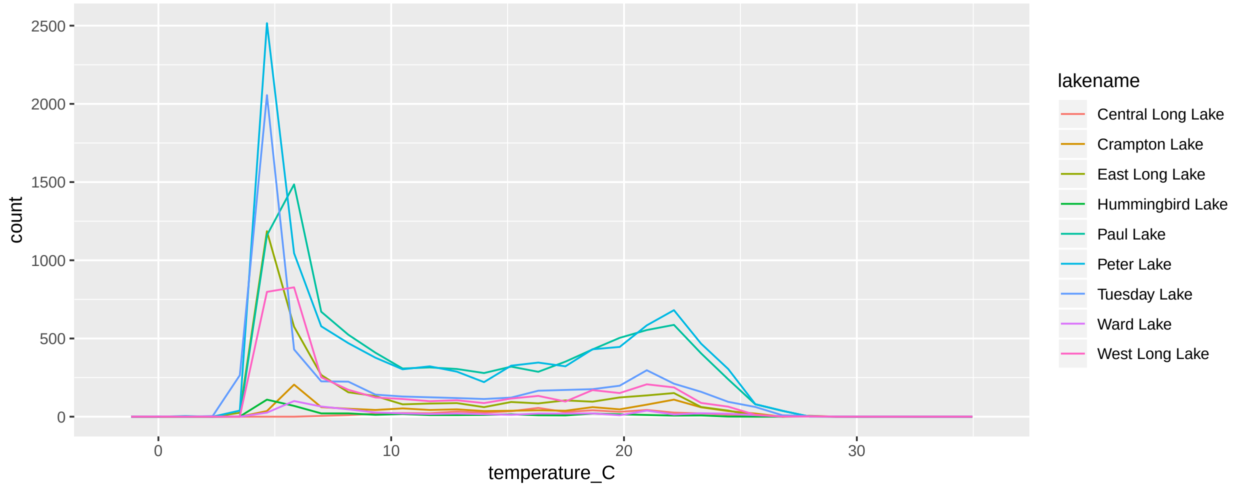


```
# 3 Change histogram from 2 to have a different number or width of bins

ggplot(temp_L_raw) +
  geom_histogram(aes(x = temperature_C), binwidth = 2)
```

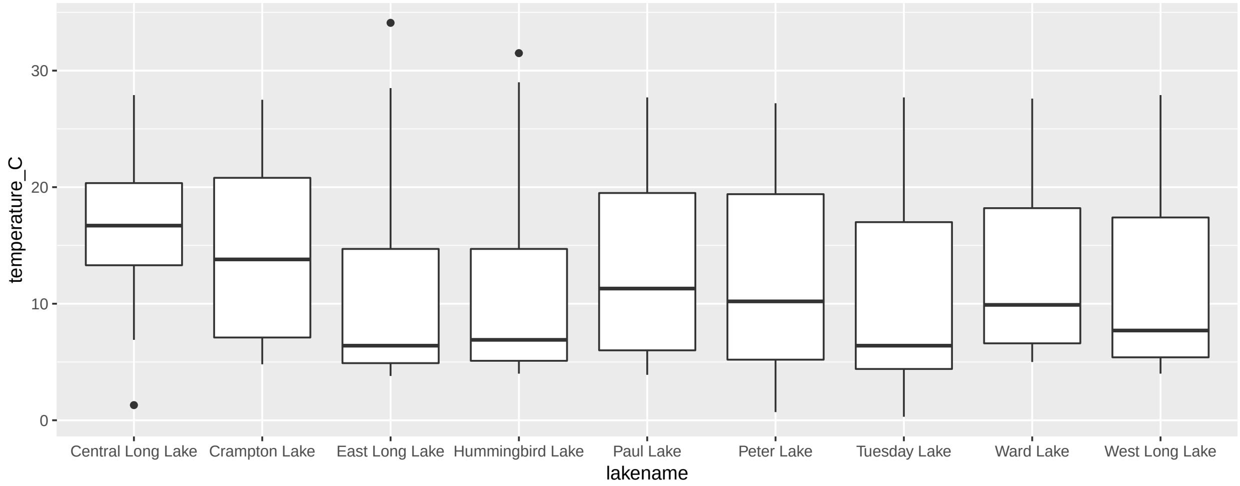


```
# 4 Frequency polygon of temperature for each lake. Choose different colors for each lake.
ggplot(temp_L_raw,aes(x = temperature_C)) +
  geom_freqpoly(aes(color = lakename))
```

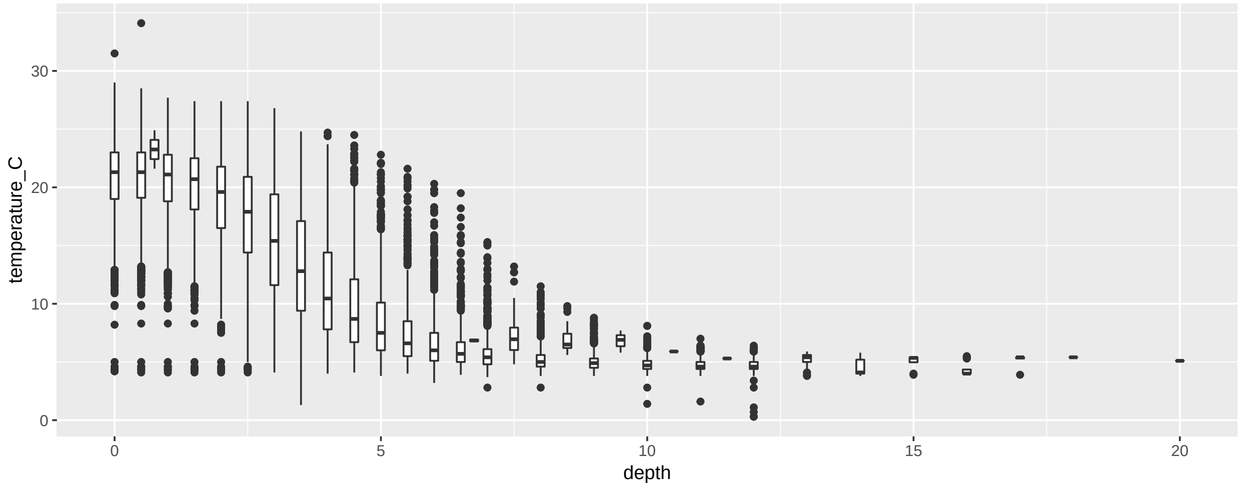
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



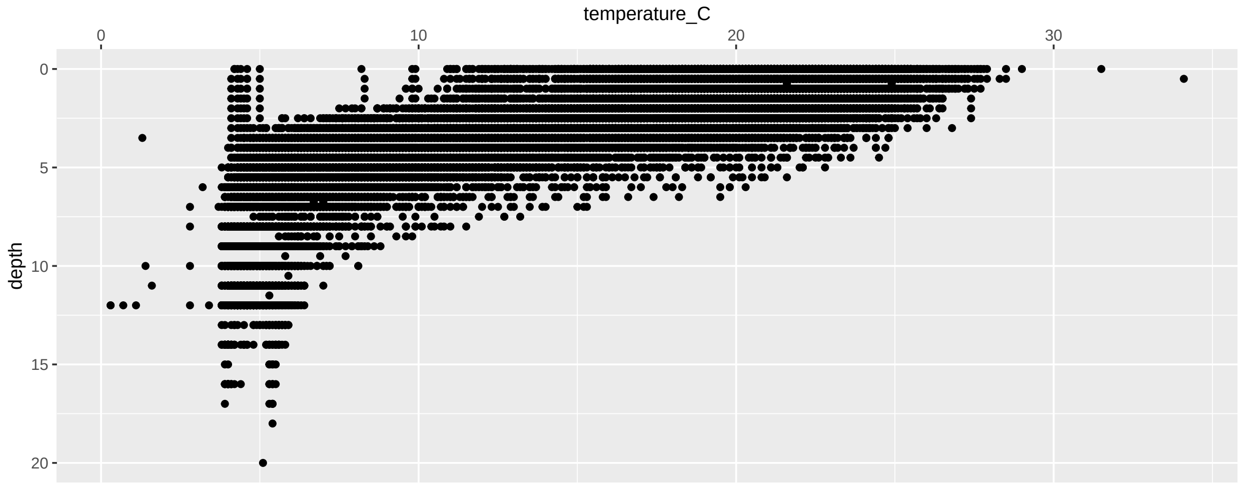
```r
# 5 Boxplot of temperature for each lake

ggplot(temp_L_raw) +
  geom_boxplot(aes(x = lakename, y= temperature_C), position = "dodge" )
```



```r
# 6 Boxplot of temperature based on depth, with depth divided into 0.25 m increments

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



```
# 7 Scatterplot of temperature by depth

P<- ggplot(temp_L_raw) +
  geom_point(aes(x = temperature_C, y = depth))

P+ scale_y_reverse() + scale_x_continuous(position = 'top')
```



## 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 out that the sampling frequency of all the temperate lakes highly varied over the sampling period. I also observed that the chemical nd physical parameters had a high amount of NAs, which definitely had to be accounted for. Most of the lakes had a distinct thermocline and showed that with increasing depth, the temperature decreased. Additionally, the temperature range in celsius between all the lakes was 0.30 C - 34.10 C.

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

> ANSWER 1: How is dissolved oxygen and temperature affected by the seasons?

> ANSWER 2: Is there a significant difference between the different temperate lakes in regards to water temperature?

ANSWER 3: Is there a correlation between depth and dissolved oxygen concentration?