Assignment 3: Data Exploration

Ying Wei Jong

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.

```
setwd("/Users/YwJong/Documents/NSOE/Spring 2019/ENV 872 Environment Data Analytics/Labs")
library(tidyverse)
```

```
## -- 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
```

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

```
LTER <- read.csv("/Users/YwJong/Documents/NSOE/Spring 2019/ENV 872 Environment Data Analytics/Labs/Data/
```

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 LTER data was divided into three different csvs. One focuses on the Carbon content, one on Nutrient content and another on physical and chemical limnology properties of Lakes at Cascade Project at North Temperate Lakes. We are looking at the csv that has physical and chemical limnology properties. This csv contains information of physical and chemical readings at various depths measured using methods described by Carpenter & Kitchell (1993) and Carpenter et al (2001).

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 dismensions of the dataset
dim(LTER)
```

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

```
# 2 class of the dataset
class(LTER)
```

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

```
# 3
head(LTER,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
str(LTER)
```

```
## 'data.frame':    38614 obs. of  11 variables:
##  $ lakeid         : Factor w/ 9 levels "C","E","H","L",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ lakename       : Factor w/ 9 levels "Central Long Lake",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ year4          : int  1984 1984 1984 1984 1984 1984 1984 1984 1984 1984 ...
##  $ daynum         : int  148 148 148 148 148 148 148 148 148 148 ...
```

```
##  $ sampledate     : Factor w/ 1712 levels "10/1/07","10/1/93",..: 134 134 134 134 134 134 134 134 134
##  $ depth          : num  0 0.25 0.5 0.75 1 1.5 2 3 4 5 ...
##  $ temperature_C  : num  14.5 NA NA NA 14.5 NA 14.2 11 7 6.1 ...
##  $ dissolvedOxygen: num  9.5 NA NA NA 8.8 NA 8.6 11.5 11.9 2.5 ...
##  $ irradianceWater: num  1750 1550 1150 975 870 610 420 220 100 34 ...
##  $ irradianceDeck : num  1620 1620 1620 1620 1620 1620 1620 1620 1620 1620 ...
##  $ comments       : Factor w/ 2 levels "DO Probe bad - Doesn't go to zero",..: NA NA NA NA NA NA NA
```

```
# 5
summary(LTER$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(LTER$depth)
```

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

```
summary(LTER$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 sampledate to class=date
LTER$sampledate <- as.Date(LTER$sampledate, format = "%m/%d/%y")
#Display class of LTER$sampledate
str(LTER)
```

```
## 'data.frame':    38614 obs. of  11 variables:
##  $ lakeid         : Factor w/ 9 levels "C","E","H","L",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ lakename       : Factor w/ 9 levels "Central Long Lake",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ year4          : int  1984 1984 1984 1984 1984 1984 1984 1984 1984 1984 ...
##  $ daynum         : int  148 148 148 148 148 148 148 148 148 148 ...
##  $ sampledate     : Date, format: "1984-05-27" "1984-05-27" ...
##  $ depth          : num  0 0.25 0.5 0.75 1 1.5 2 3 4 5 ...
##  $ temperature_C  : num  14.5 NA NA NA 14.5 NA 14.2 11 7 6.1 ...
##  $ dissolvedOxygen: num  9.5 NA NA NA 8.8 NA 8.6 11.5 11.9 2.5 ...
##  $ irradianceWater: num  1750 1550 1150 975 870 610 420 220 100 34 ...
##  $ irradianceDeck : num  1620 1620 1620 1620 1620 1620 1620 1620 1620 1620 ...
##  $ comments       : Factor w/ 2 levels "DO Probe bad - Doesn't go to zero",..: NA NA NA NA NA NA NA
```

```
#first 10 rows of date column
head(LTER$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: I tend not to remove NAs until I absolutely have to. For example, I might keep all the NAs in temperature until when I run a regression and the model does not accept any NAs.

I will also keep all the NAs unless the NA was in a column so critical in analysis that the data point was unusable anyway.

4) Explore your data graphically

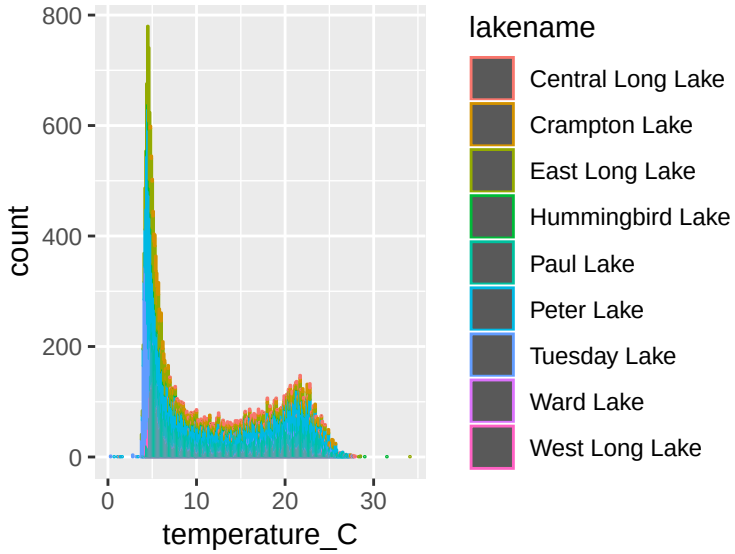
Write R commands to display graphs depicting:

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. I personally don't think this is the best way to dis
ggplot(LTER) +
  geom_bar(aes(x=temperature_C, color=lakename), position = "stack")
```

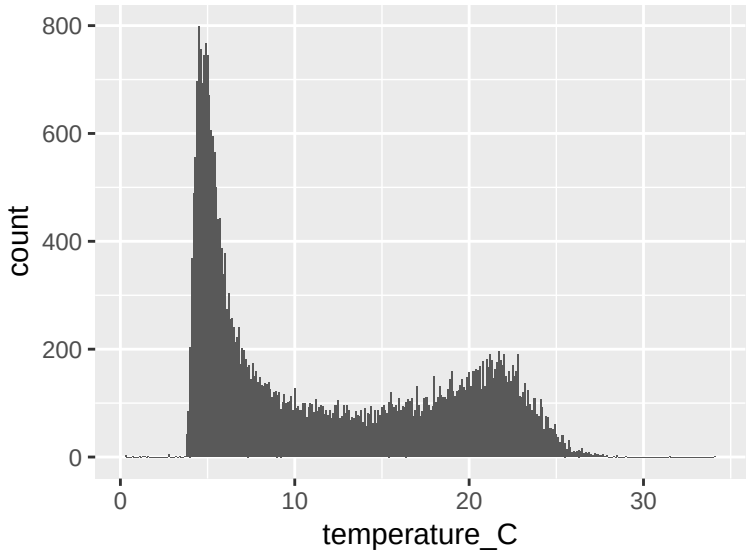
```
## Warning: Removed 3858 rows containing non-finite values (stat_count).
```

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



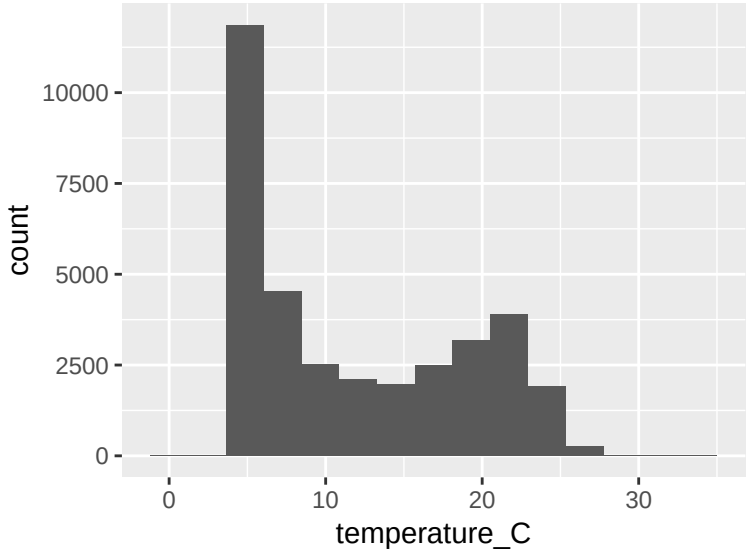

```
# 2 Histogram of count distributions of temperature (all temp measurements together)
ggplot(LTER) +
  geom_histogram(aes(x=temperature_C), binwidth = 0.1)
```

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



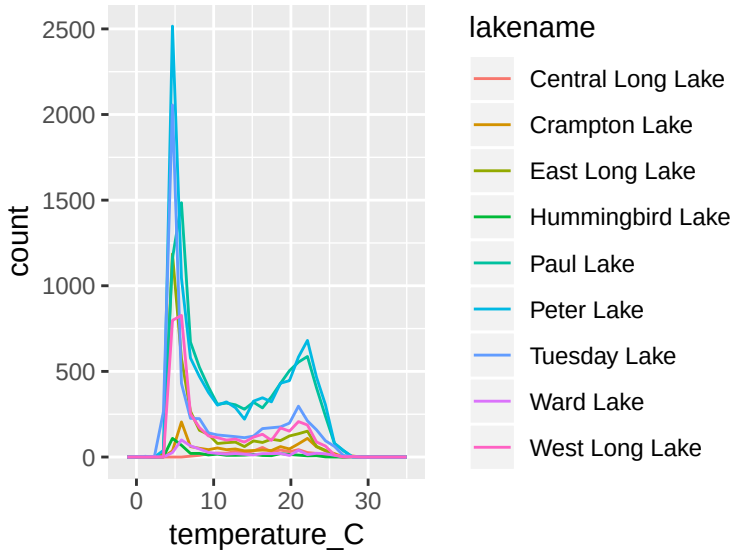

```
# 3 Histogram of count distributions of temperature with lower number of bins, i.e. lower resolution
ggplot(LTER) +
  geom_histogram(aes(x=temperature_C), bins = 15)
```

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



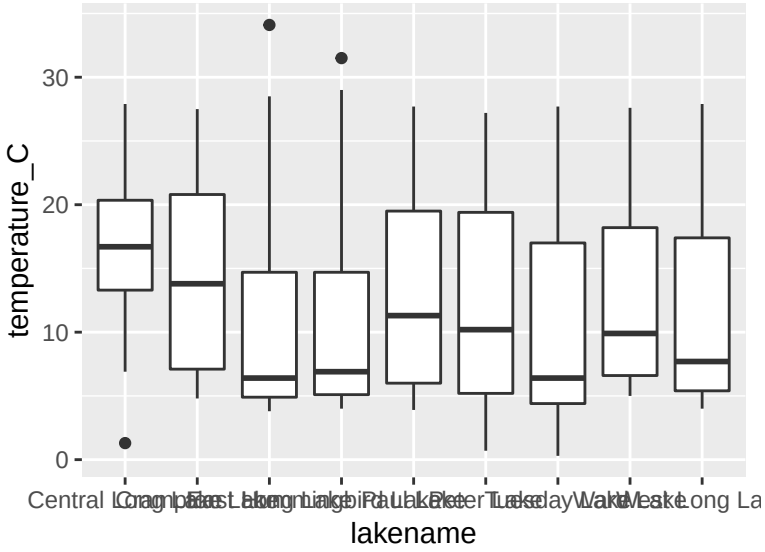

```
# 4 Frequency polygon of temperature for each lake. Different colors for each lake.
ggplot(LTER) +
  geom_freqpoly(aes(x=temperature_C, color=lakename))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 3858 rows containing non-finite values (stat_bin).
```



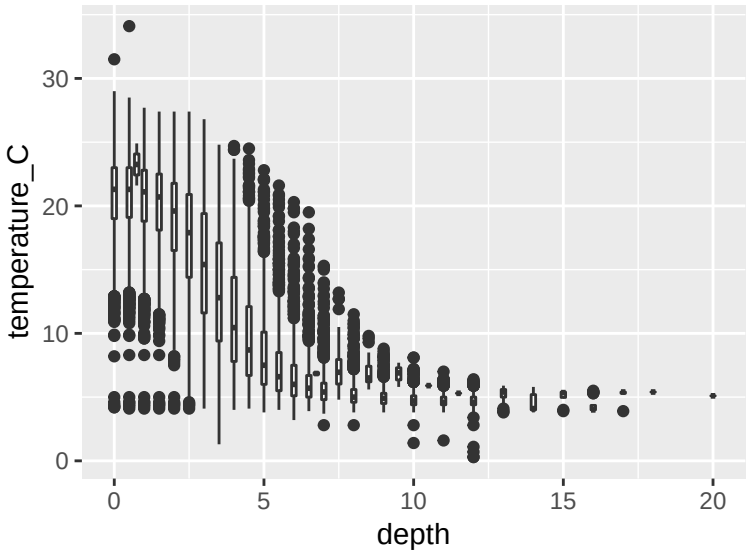

```
# 5 Boxplot of temperature for each lake
ggplot(LTER) +
  geom_boxplot(aes(x = lakename, y = temperature_C))
```

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



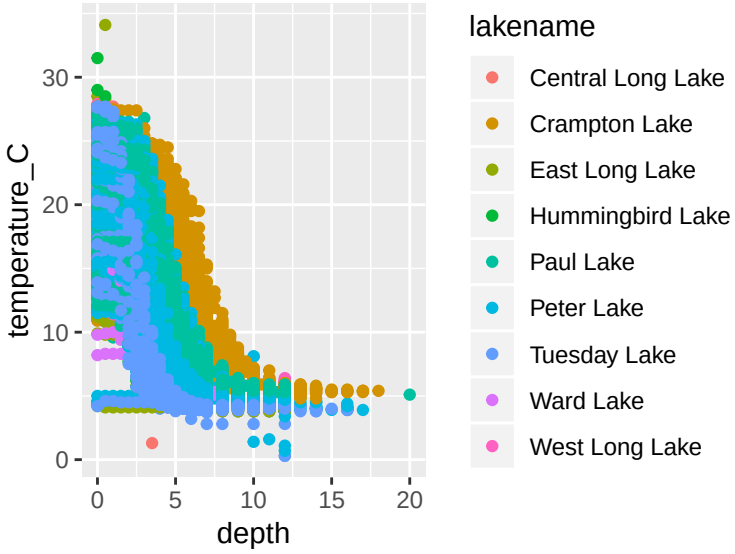

```
# 6 Boxplot of temperature based on depth, with depth divided into 0.25 m increments
ggplot(LTER) +
  geom_boxplot(aes(x = depth, y = temperature_C, group=cut_width(depth, 0.25)))
```

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




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

```
## 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: Some lakes are more represented than others in this dataset (i.e. some lakes have more data than others). Each lake has different median temperature. Most temperatures observations were about 4 celsius degrees. The deeper the lake depth, the lower the temperature.

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

> ANSWER 1: How does the average temperature change throughout the years?
>
> ANSWER 2: How do the dissolved oxygen and irradiance reading change with lake depth?

ANSWER 3: Are there any correlation between depth and dissolved oxygen (and between pairs of other variables)?