Assignment 3: Data Exploration

Wanchen Xiong

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 inst alled 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] "C:/Users/Wanch/Desktop/ENVI 872 data/Environmental_Data_Analytics/Assignments"
```

```
setwd("C:/Users/Wanch/Desktop/ENVI 872 data/Environmental_Data_Analytics")

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------- tidyverse 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_confli
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
monitoring_threelakes <- 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: This file contains data from studies on several lakes in the North Temperate Lakes District in Wisconsin, USA. Data were collected as part of the Long Term Ecological Research station established by the National Science Foundation. The three aspects of the data are Carbon, which covers from 1984 to 2016, nutrients, which covers from 1991 to 2016, and physical and chemical limnology, which covers from 1984 to 2016.

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 dimension of the dataset:

dim(monitoring_threelakes)
```

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

```
# 2 Display the class of the dataset:

class(monitoring_threelakes)
```

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

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

head(monitoring_threelakes)
```

```
##   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
##   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>
```

```
# 4 Display the class of each variable: lakename, sampledata, depth, and temperature

class(monitoring_threelakes$lakename)
```

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

```
class(monitoring_threelakes$sampledate)
```

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

```
class(monitoring_threelakes$depth)
```

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

```
class(monitoring_threelakes$temperature_C)
```

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

```
# 5 Summarize the data under lakename, depth, and temperature

summary(monitoring_threelakes$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(monitoring_threelakes$depth)
```

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

```
summary(monitoring_threelakes$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.

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

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

```
head(monitoring_threelakes$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: It is ok to remove the NAs because they just stood for missing data.

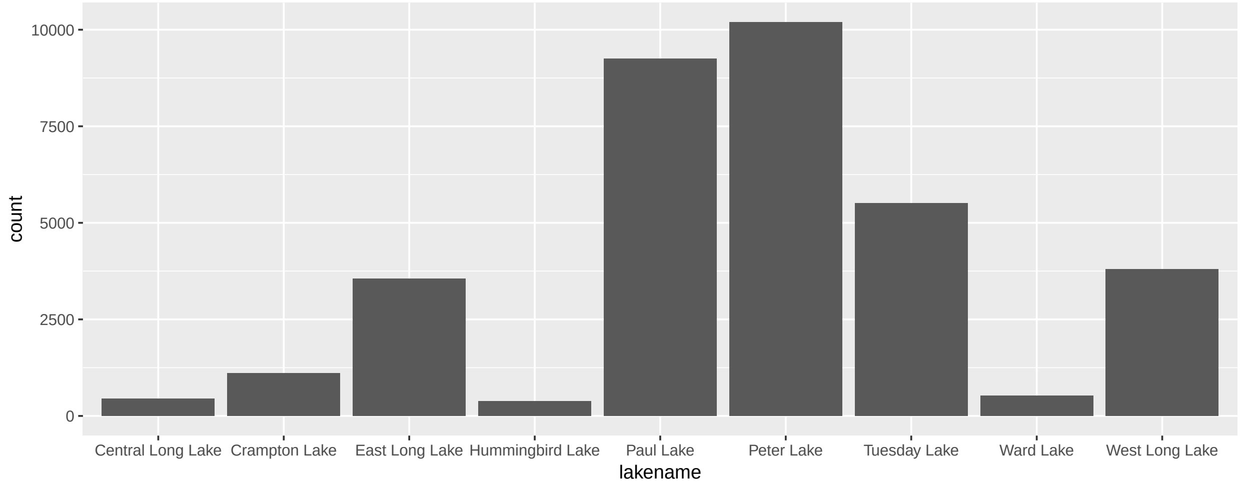
4) Explore your data graphically

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

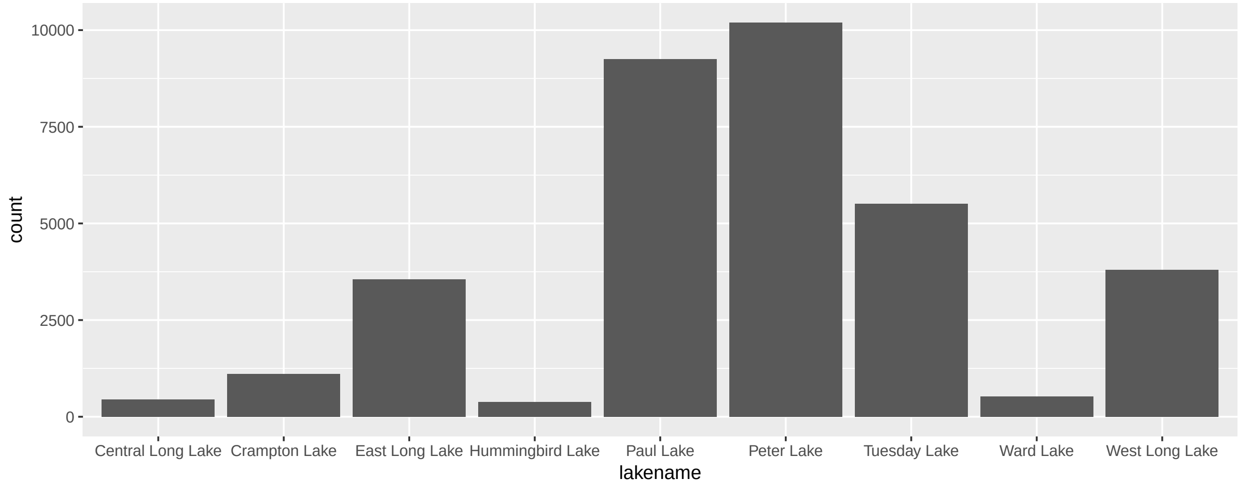
```
threelakes_temp <- select(monitoring_threelakes, "lakename", "temperature_C")
threelakes_complete <- na.omit(threelakes_temp)

# 1 Create a bar chart of temperature counts for each lake
```

```
ggplot(threelakes_complete, aes(x = lakename)) +
 geom_bar()
```



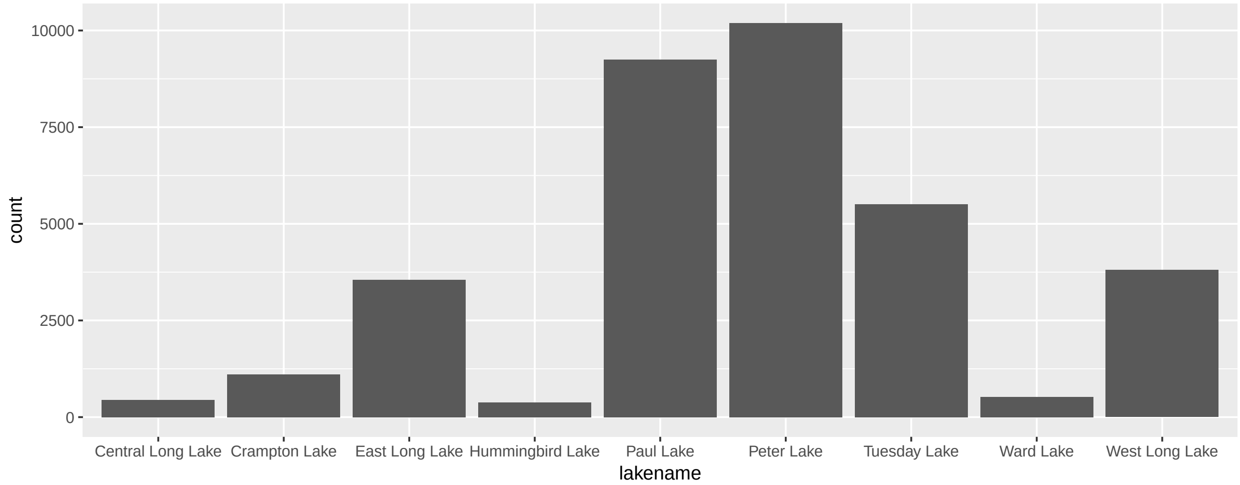
```
# 2 Create a histogram of count distributions of temperature (all temp measurements together)
ggplot(threelakes_complete) +
 geom_histogram(aes(x = lakename), stat = "count")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



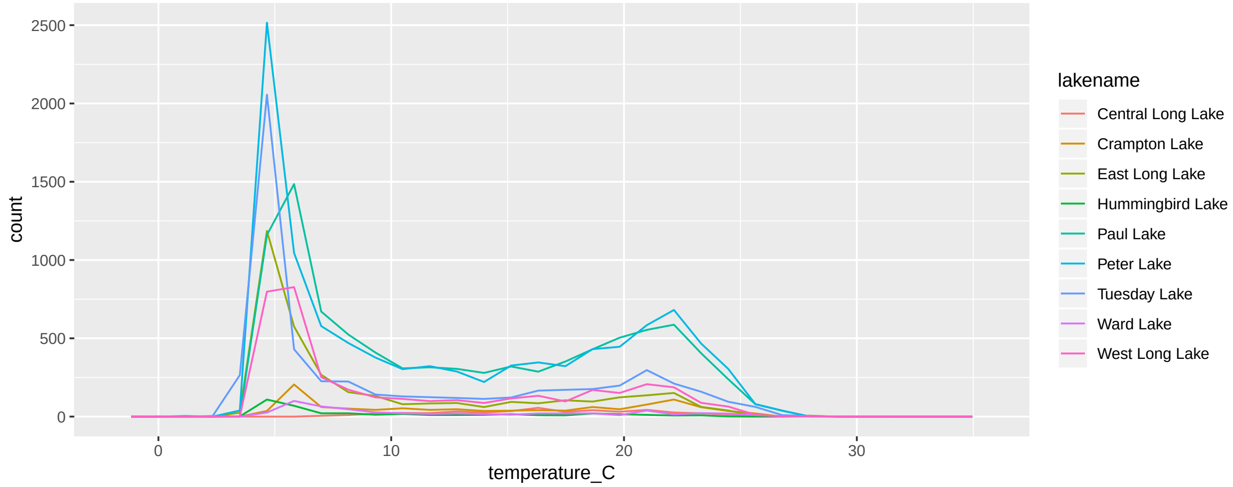
```
# 3 Change histogram from 2 to have a different number or width of bins
ggplot(threelakes_complete) +
  geom_histogram(aes(x = lakename), stat = "count",
                 bins = 60 )
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

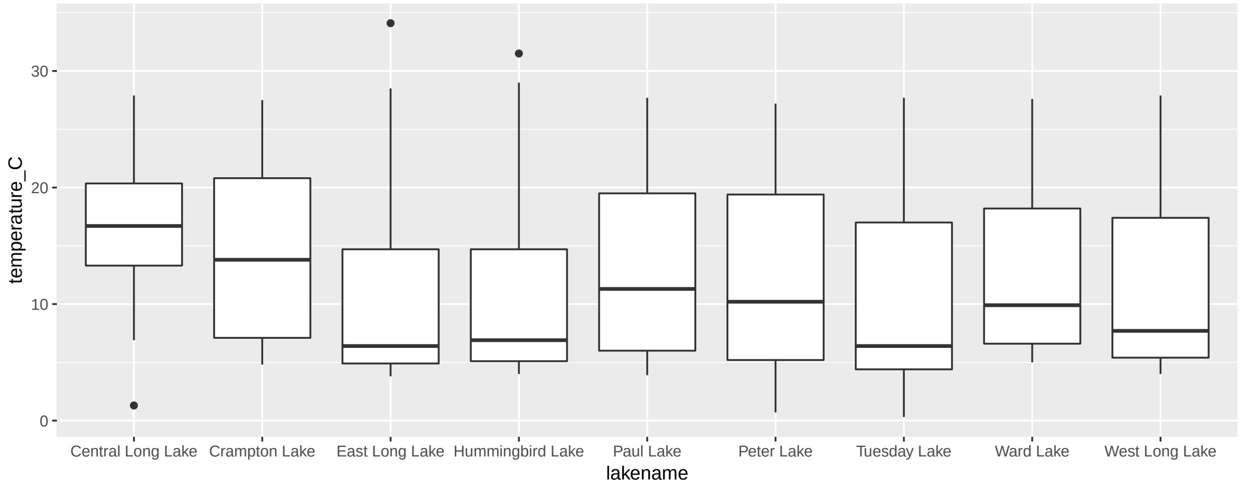


```
# 4 Create a frequency polygon of temperature for each lake. Choose different colors for each lake.
ggplot(threelakes_complete) +
  geom_freqpoly(aes(x = temperature_C, color = lakename))
```

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

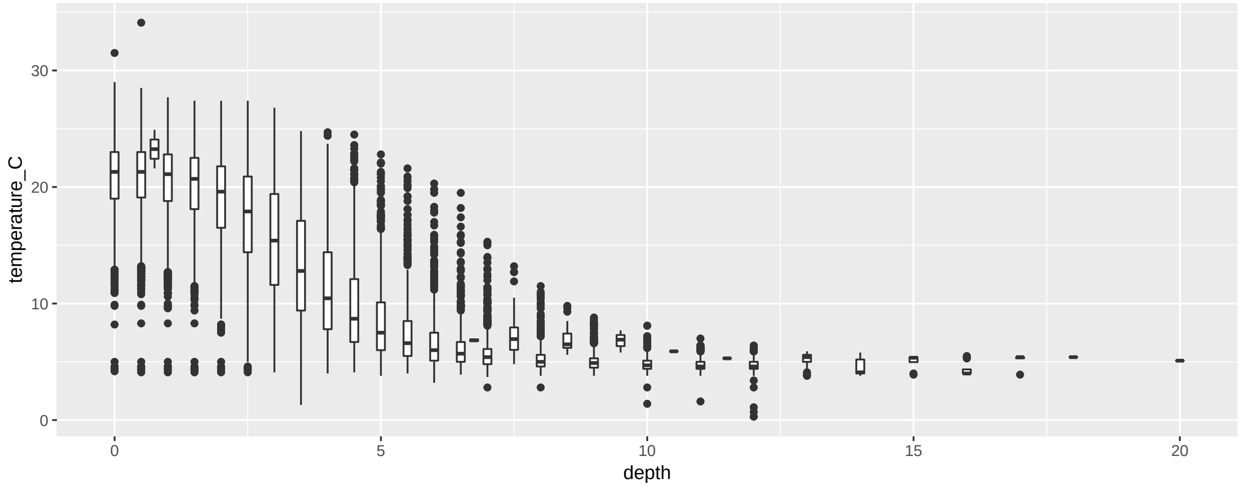


```
# 5 Create a boxplot of temperature for each lake
ggplot(threelakes_complete) +
  geom_boxplot(aes(x = lakename, y = temperature_C))
```



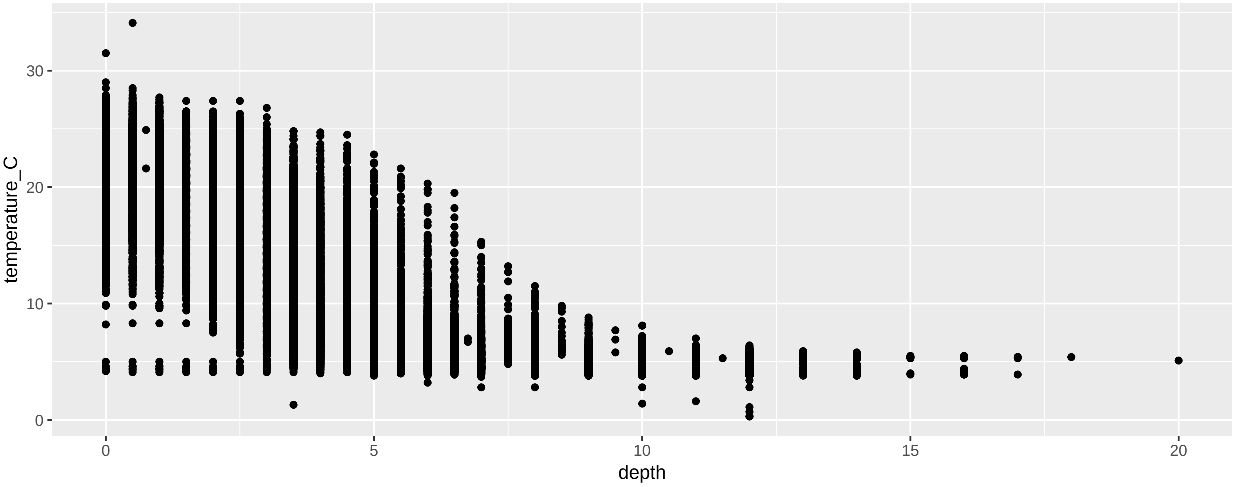
```
# 6 Create a boxplot of temperature based on depth, with depth divided into 0.25 m increments
ggplot(monitoring_threelakes) +
  geom_boxplot(aes(x = depth, y = temperature_C, group = cut_width(depth, 0.25)))
```

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



```
# 7 Create a scatterplot of temperature by depth
ggplot(monitoring_threelakes) +
  geom_point(aes(x = depth, 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: According to the histogram, Paul lake and Peter lake are two lakes that have the most temperature data. Across the nine lakes, shown by the frequency graph, the highest frequency of temperature is around 4-5 degree Celsius. According to the box plot, nine lakes have various average temperature, ranging from 6-7 degree Celsius to 17-18 degree Celsius. And the scatterplot reflects 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: Do all the lakes have similar pattern of the relationship between temperature and depth?
>
> ANSWER 2: What's the relationship between depth and dissolved oxygen?
>
> ANSWER 3: What's the relationship between temperature and dissolved oxygen?