Assignment 3: Data Exploration

Felipe Raby Amadori

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.

```
# working directory should be the parent folder for the Environmental Data Analytics Course
# this specific file path only works in Felipe's Computer
setwd("C:/Users/Felipe/OneDrive - Duke University/1. DUKE/1. Ramos 2 Semestre/EOS-872 Env. Data Analyti
# Load package
library(tidyverse)
```

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

```
## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     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()
```

```
# Relative file path (friendly for users regardless of machine)
NTL_LTER_Lake.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- What are the data contained in the file. 2- Where were the data collected 3- How were the data collected. Also you know where to reach out in case of doubts.

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_LTER_Lake.data)
```

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

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

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

```
# 3
head(NTL_LTER_Lake.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_LTER_Lake.data$lakename)
```

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

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

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

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

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

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

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

```
# 5
summary(NTL_LTER_Lake.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_LTER_Lake.data$depth)
```

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

```
summary(NTL_LTER_Lake.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 sampledate is indeed date. Write another R command to show the first 10 rows of the date column.

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

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

```
head(NTL_LTER_Lake.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: We do not know yet the research question so we do not know if we want to remove NAs from this dataset. also we need to be careful because there is a comment column with mostly NA values, which doesn't mean that the row is invalid.

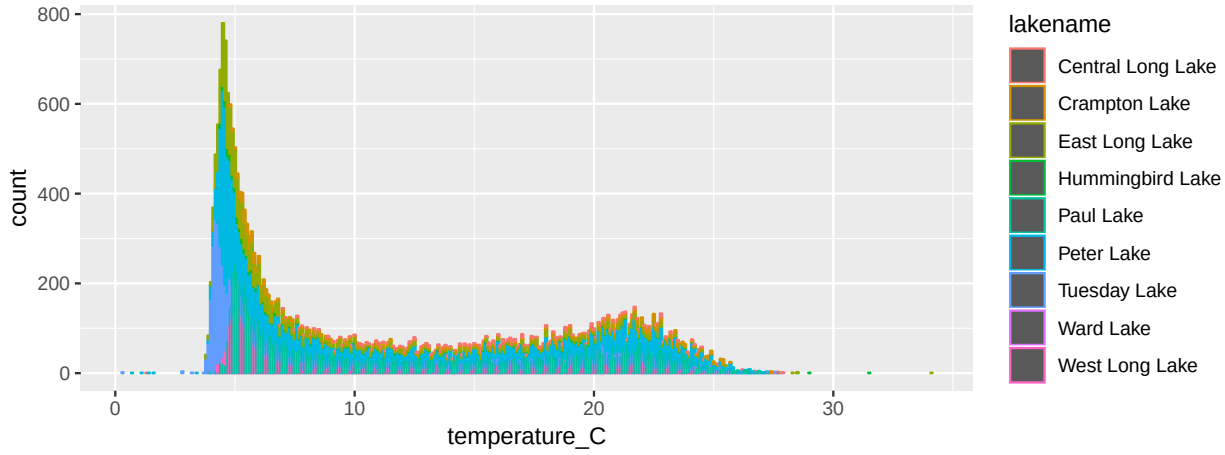
4) Explore your data graphically

Write R commands to display graphs depicting:

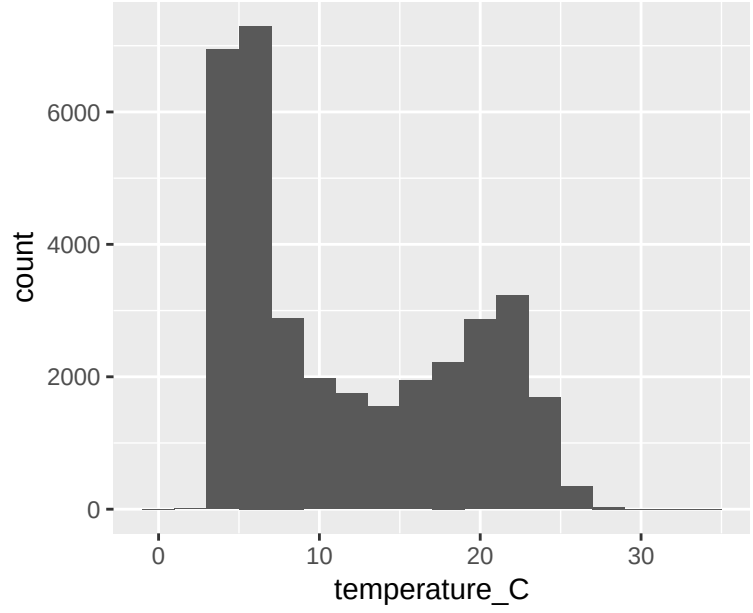
1. Bar chart of temperature counts for each lake
2. Histogram of count distributions of temperature (all temp measurements together)
3. Change histogram from 2 to have a different number or width of bins
4. Frequency polygon of temperature for each lake. Choose different colors for each lake.
5. Boxplot 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
# Rows with Temperature NAs are removed because in a temperature variable NAs are not useful.
# Only rows with NAs in temperature where deleted.

NTL_LTER_Lake.data.complete <- subset(NTL_LTER_Lake.data, !is.na(temperature_C))

ggplot(NTL_LTER_Lake.data.complete, aes(x = temperature_C, color = lakename)) +
  geom_bar()
```

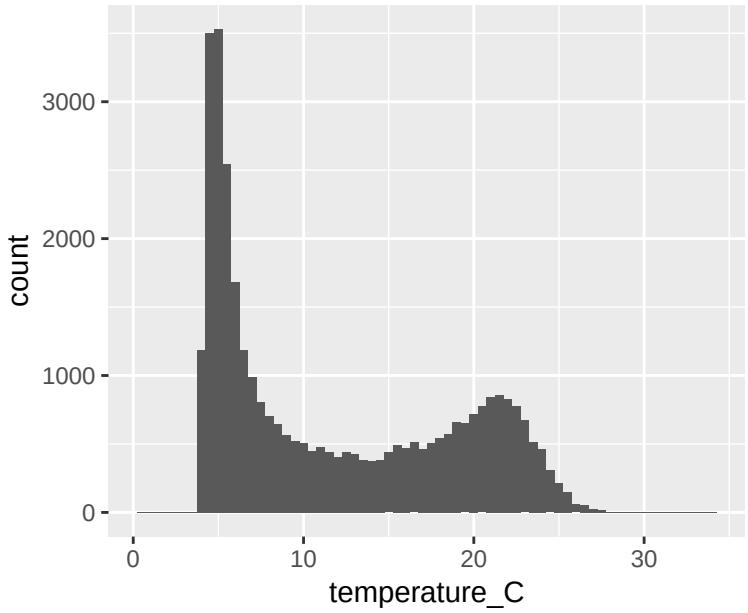
```
## Warning: position_stack requires non-overlapping x intervals
```



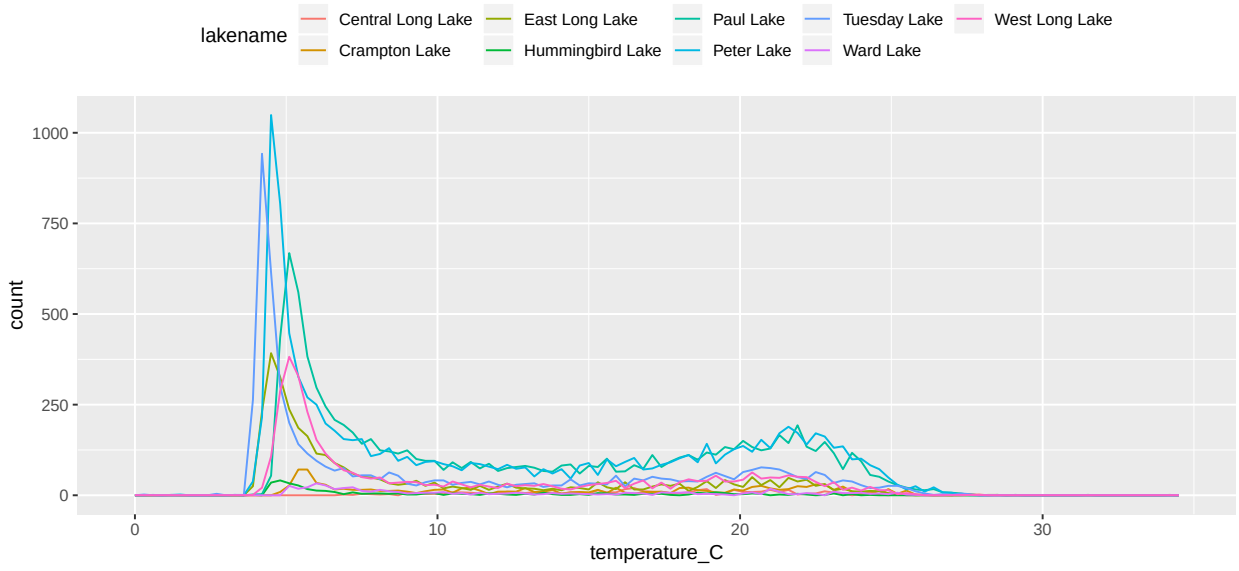

```
# 2
ggplot(NTL_LTER_Lake.data.complete) +
  geom_histogram(aes(x = temperature_C), binwidth = 2)
```



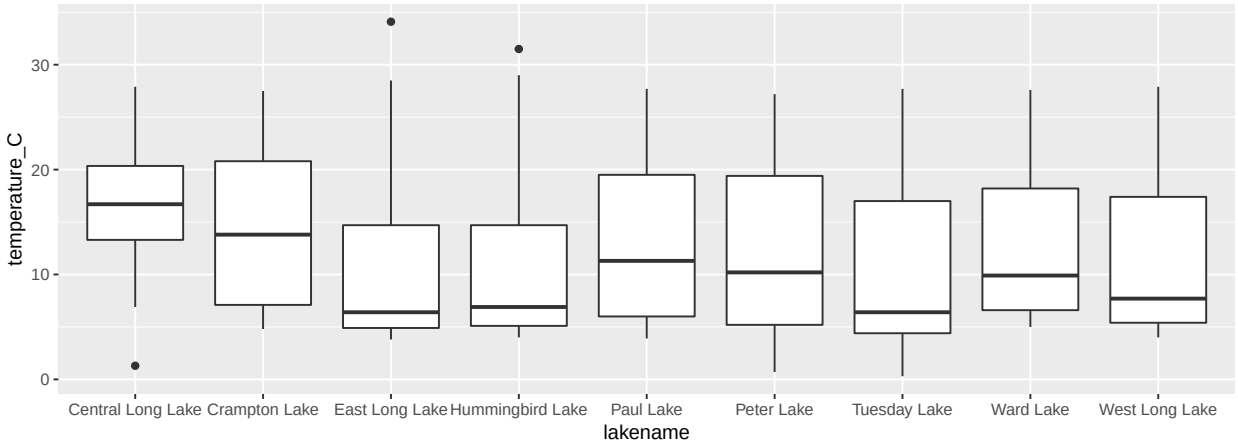

```
# 3
ggplot(NTL_LTER_Lake.data.complete) +
  geom_histogram(aes(x = temperature_C), binwidth = 0.5)
```



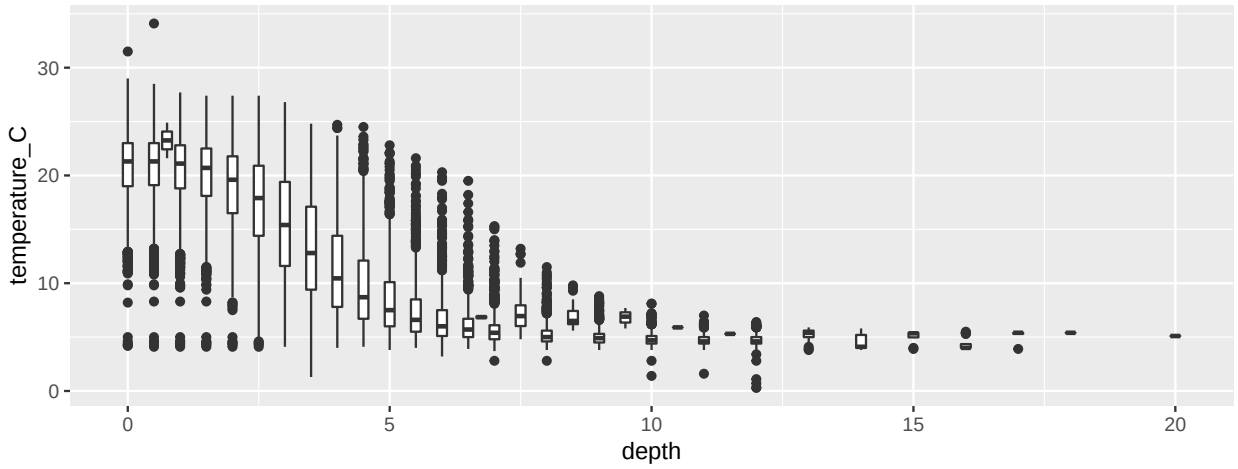

```
# 4
ggplot(NTL_LTER_Lake.data.complete) +
  geom_freqpoly(aes(x = temperature_C, color = lakename), binwidth = 0.3) +
  theme(legend.position = "top")
```



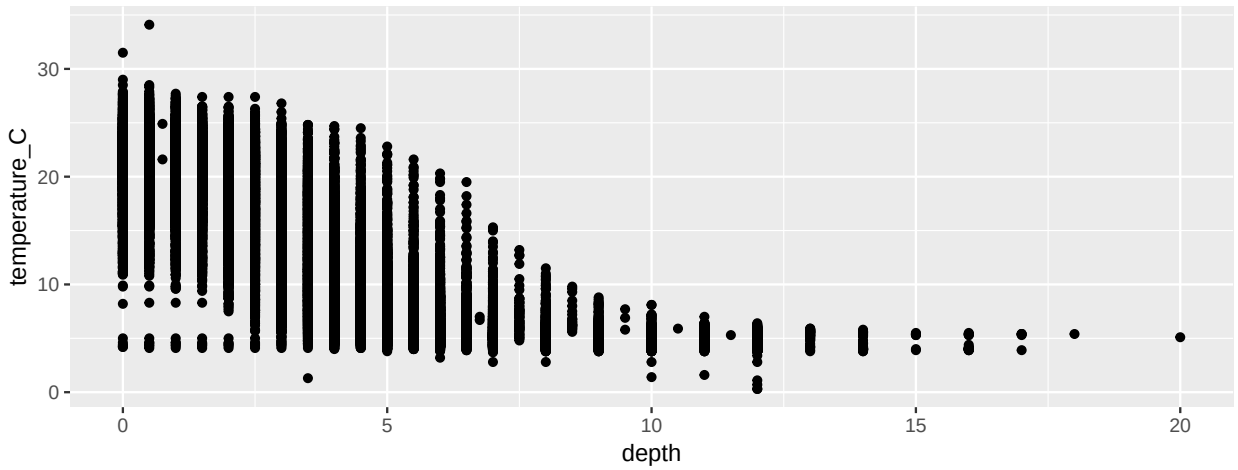

```
# 5
ggplot(NTL_LTER_Lake.data.complete) +
  geom_boxplot(aes(x = lakename, y = temperature_C))
```




```
# 6
ggplot(NTL_LTER_Lake.data.complete) +
  geom_boxplot(aes(x = depth, y = temperature_C, group = cut_width(depth, 0.25)))
```




```
# 7
ggplot(NTL_LTER_Lake.data.complete) +
  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 out the variables in the data set, the name of the lakes involved in the study, the statistics of the variables depth and temperature. Also the bar chart, histogram, frequency polygons, and the first boxplot showed me the frequency and range of temperature values globally and for each lake. Finally the boxplot and scatterplot using temperature and depth showed me the relationship between depth and temperature. Temperatures tend to decrease with depth.

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

> ANSWER 1: What is the relationship between the other variables in the data set?
>
> ANSWER 2: Does the relationship between Temperature and Depth the same in each lake?
>
> ANSWER 3: How strong is the relationship between Temperature and Depth implied by the scatterplot and the boxplot?