Assignment 3: Data Exploration

Kat Horan

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.

```
#set working directory
#setwd("/Users/kathleenhoran/Desktop/Duke/Spring 2019/Env. Data Analytics/Env_Data_Analytics")
getwd()
```

```
## [1] "/Users/kathleenhoran/Desktop/Duke/Spring 2019/Env. Data Analytics/Env_Data_Analytics/Assignments
```

```
#load packages
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------- tidyverse 1
```

```
## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.7
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

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

```
#upload data
NTL.LTER.ChemPhys <- read.csv("/Users/kathleenhoran/Desktop/Duke/Spring 2019/Env. Data Analytics/Env_Da
```

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 dataset 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.

> ANSWER: 2.) Inorganic samples were collected at depths corresponding to 100%, 50%, 25%, 10%, 5%, and 1% of surface irradiance, as well as one sample from the hypolimnion.

> ANSWER: 3.) Files are named according to the following naming convention: `databasename_datatype_details_stage.`

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

```r
# 1 - Dimensions of the dataset
dim(NTL.LTER.ChemPhys)
```

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

```r
length(NTL.LTER.ChemPhys)
```

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

```r
# 2 - Class of the dataset
class(NTL.LTER.ChemPhys)
```

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

```r
# 3 - First 8 rows of the dataset
head(NTL.LTER.ChemPhys, 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>
```

```r
# 4 - Class of the variables
class(NTL.LTER.ChemPhys$lakename) #factor
```

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

```r
class(NTL.LTER.ChemPhys$sampledate) #factor
```

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

```r
class(NTL.LTER.ChemPhys$depth) #numeric
```

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

```r
class(NTL.LTER.ChemPhys$temperature_C) #numeric
```

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

```r
# 5 - Summary of variables
summary(NTL.LTER.ChemPhys$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(NTL.LTER.ChemPhys$depth)
```

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

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

#Confirm class has been changed
class(NTL.LTER.ChemPhys$sampledate)
```

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

```r
#Show first 10 rows of the date column
head(NTL.LTER.ChemPhys$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 think this would depend on what we plan to be testing. Overall I would think that we would not want to remove the NAs from this dataset. Just because a sample is missing information for one of the variables does not mean that the rest of the information isn't valuable. If we are testing something relating to a specific variable that has N/A in it and we only want to

consider the samples we have values for, then it may be helpful to set up a 2nd dataframe with the N/A's removed in that variable's column. For this data set, if we used the code discussed in class (e.g. NTL.LTER.ChemPhys.complete <- na.omit(NTL.LTER.ChemPhys)) it would remove every row that has an "N/A" in the comments section which is not what we want.

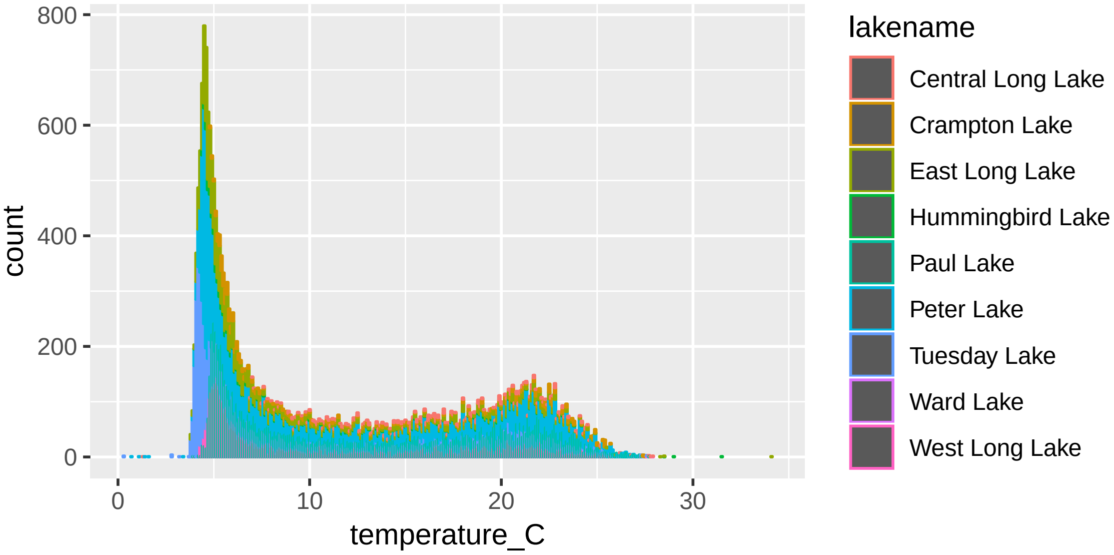
4) Explore your data graphically

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

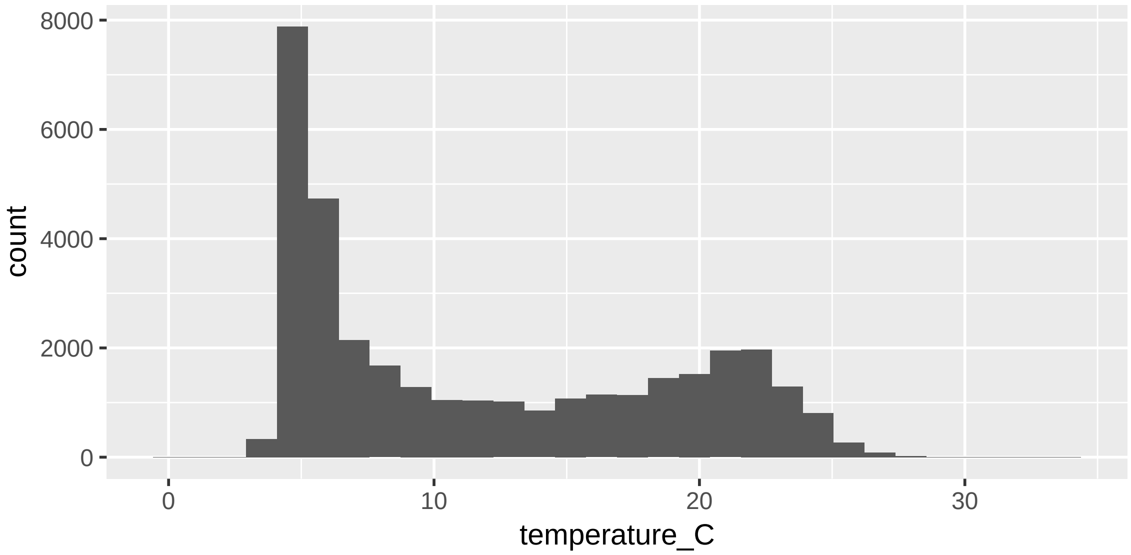
```
# 1 Bar chart
# Since we are focused on temp, I will remove the N/As from it first:
NTL.LTER.ChemPhys.TempComplete <- subset(NTL.LTER.ChemPhys, !is.na(temperature_C))

ggplot(NTL.LTER.ChemPhys.TempComplete, aes(x = temperature_C, color = lakename)) +
  geom_bar()
```

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

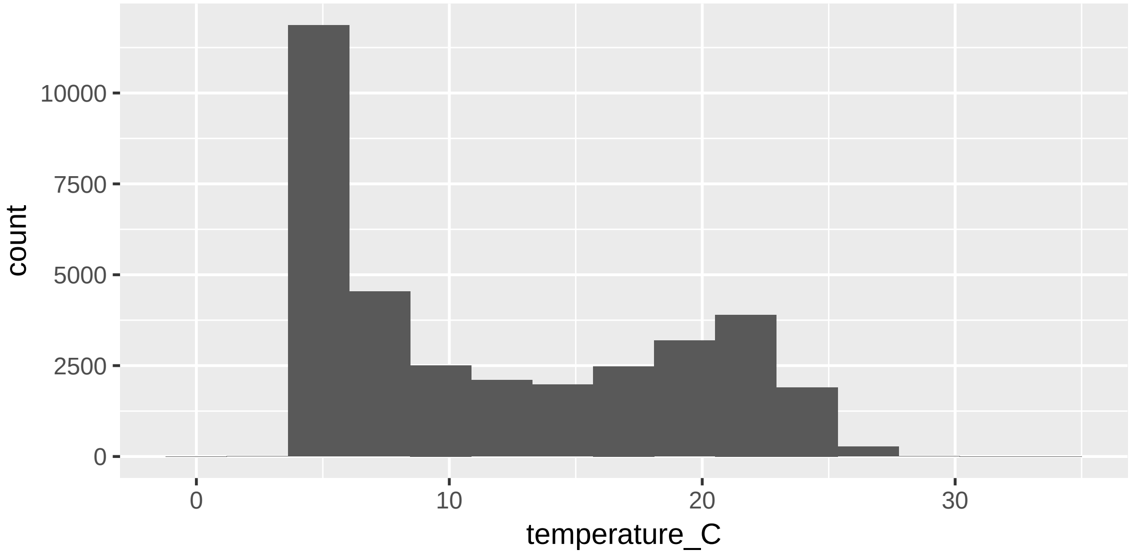


```
# 2 Histogram
ggplot(NTL.LTER.ChemPhys.TempComplete) +
  geom_histogram(aes(x = temperature_C))
```

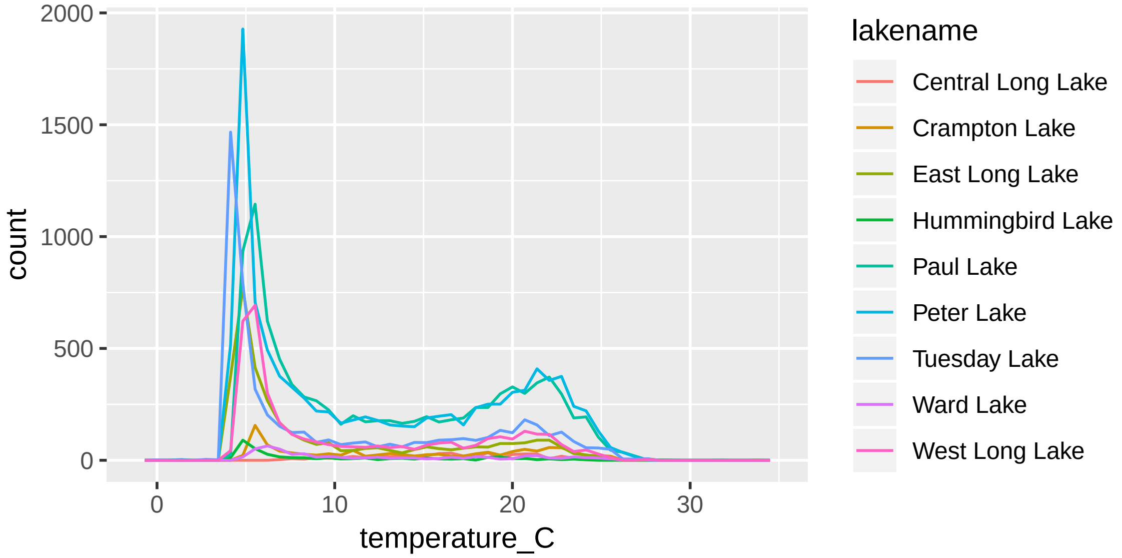
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



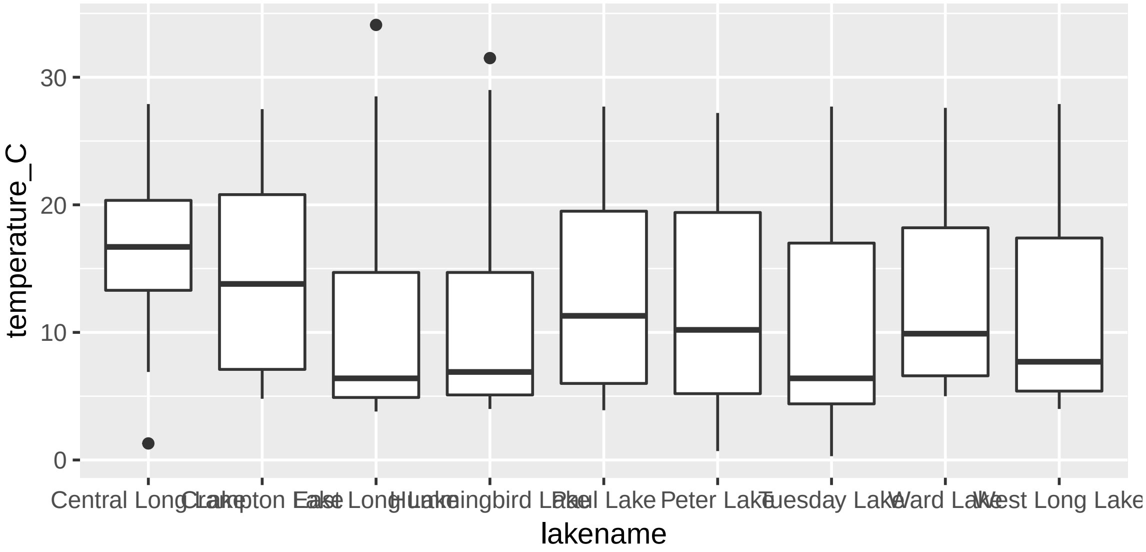
```r
# 3 Histogram, different bins
ggplot(NTL.LTER.ChemPhys.TempComplete) +
  geom_histogram(aes(x = temperature_C), bins = 15)
```



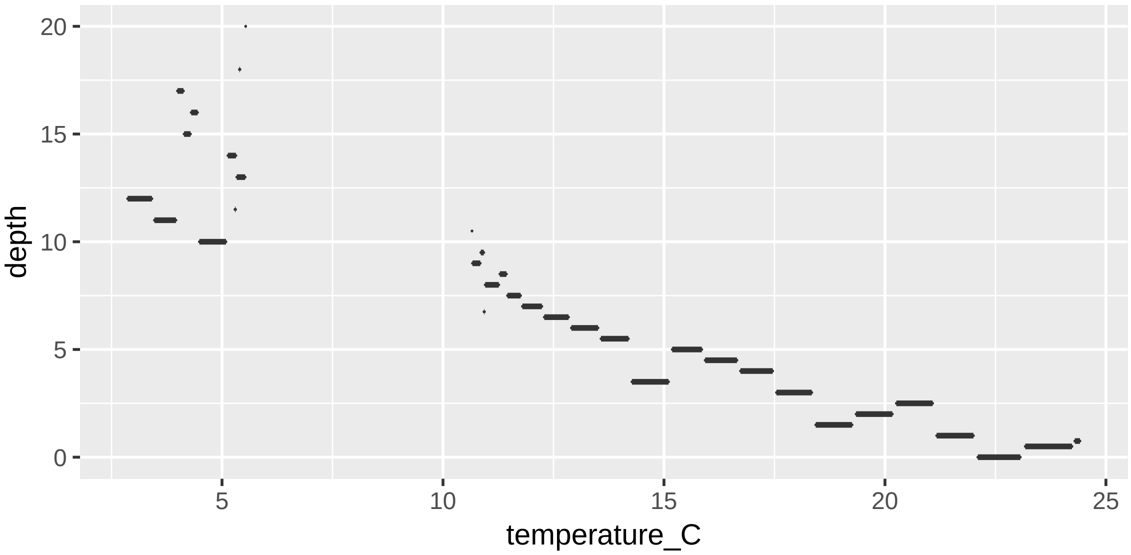
```r
# 4 Frequency polygon
ggplot(NTL.LTER.ChemPhys.TempComplete) +
  geom_freqpoly(aes(x = temperature_C, color = lakename), bins = 50)
```



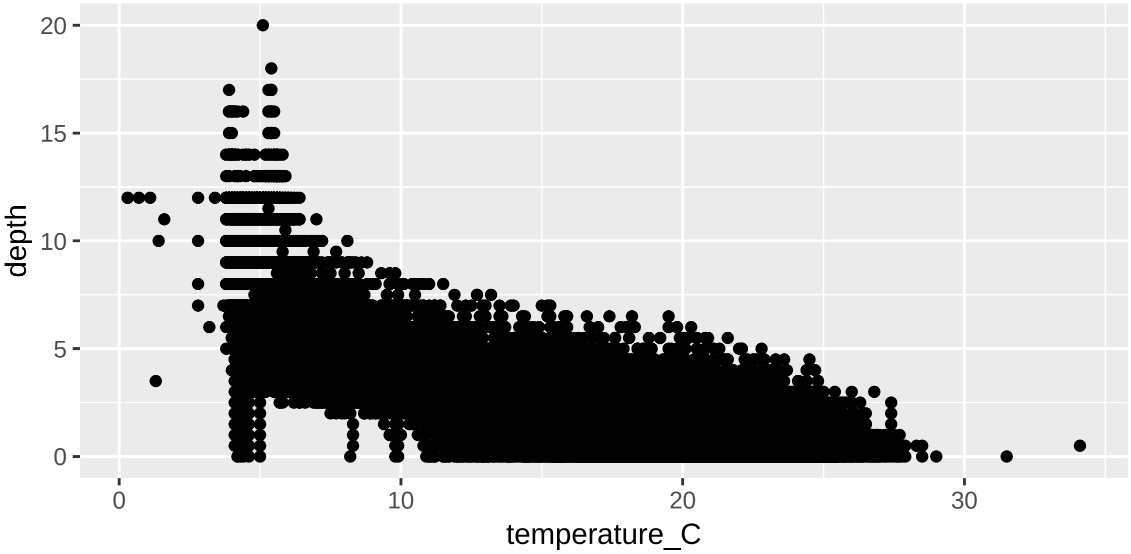
```
# 5 Boxplot of temp for each lake
ggplot(NTL.LTER.ChemPhys.TempComplete) +
  geom_boxplot(aes(x = lakename, y = temperature_C))
```



```
# 6 Boxplot of temp based on depth, with depth divided into 0.25m increments
ggplot(NTL.LTER.ChemPhys.TempComplete) +
  geom_boxplot(aes(x = temperature_C, y = depth, group = cut_width(depth, .25)))
```



```r
# 7 Scatterplot of temp by depth
ggplot(NTL.LTER.ChemPhys.TempComplete) +
  geom_point(aes(x = temperature_C, y = depth))
```



##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: From the summary and graphs made, I found that of the total of 38,614 samples from the data set, Peter Lake and Paul lake had the highest sample counts at 11,288 and 10,325 respectively. Tuesday Lake had the next highest sample count at 6107. The mean temperature recorded was 11.81 degrees celcius, with a median of 9.3. When looking at the relationship between temperature and depth, there seems to be a correlation that the temperature is lower when the depth is higher.

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

> ANSWER 1: Is there a relationship between dissolved oxygen and temperature? ANSWER 2: Has the average temperature in each lake changed over time? ANSWER 3: What is the relationship between dissolved oxygen and depth?