Assignment 3: Data Exploration

Gaby Garcia

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.

```
getwd()
```

```
## [1] "/Users/gabrielagarcia/Desktop"
```

```
setwd("~/Desktop/Environmental Data Analytics/Environmental_Data_Analytics/Data/Raw")
library(tidyverse)
```

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

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

```
LakesData <-read.csv("NTL-LTER_Lake_ChemistryPhysics_Raw.csv")  #renaming data frame to make it easier
```

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 data set consists of parameters collected from studies of lakes in the North Temperate Lakes District in Wisconsin, and was compiled using the North Temperate Lakes Long Term Ecological Research website. 2.) All of the physical and chemical variables were measured at a central station at the deepest point of each lake; the measurements were generally taken in the morning from 8-9 A.M. 3.) For the variable DOC (dissolved organic carbon), 100-300 mL of lake water from each depth was filtered through 153 um mesh to remove large zooplankton.

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

Use dim() function to determine the number of rows and the number of columns in the data frame

```
# 1
dim(LakesData)
```

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

Use class() function to determine the class attribute (data frame)

```
# 2
class(LakesData)
```

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

Use head() function to return the first 8 rows of data set

```
head(LakesData, 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>
```

Use class function to determine the class of the following variables:

```
class(LakesData$lakename)
```

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

```
class(LakesData$sampledate)
```

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

```
class(LakesData$depth)
```

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

```
class(LakesData$temperature)
```

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

Summary of Lake Name, depth, and temperature

```
summary(LakesData$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(LakesData$depth)
```

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

```
summary(LakesData$temperature)
```

```
##    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.

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

```
class(LakesData$sampledate)
```

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

```
head(LakesData$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: No, I chose to leave the NAs in the dataset. I realized that deleting the NAs would not only remove 3,858 temperature values, but would also leave the data set with just three sampled lakes remaining. It would not make sense to remove the NA's because for instance, even though there are days where irradiance was not measured at all, it could still be useful for the data scientist who will later use this data to have access to the temperature and DO variables on these days.

There are 3,858 NA's for the Temperature variable.

```
##LakesDataComplete <- na.omit(LakesData)
sum(is.na(LakesData$temperature_C))
```

```
## [1] 3858
```

There are 4,039 NA's for the Dissolved Oxygen variable.

```
sum(is.na(LakesData$dissolvedOxygen))
```

```
## [1] 4039
```

There are 14,287 NA's for the Irradiance Water variable.

```
sum(is.na(LakesData$irradianceWater))
```

```
## [1] 14287
```

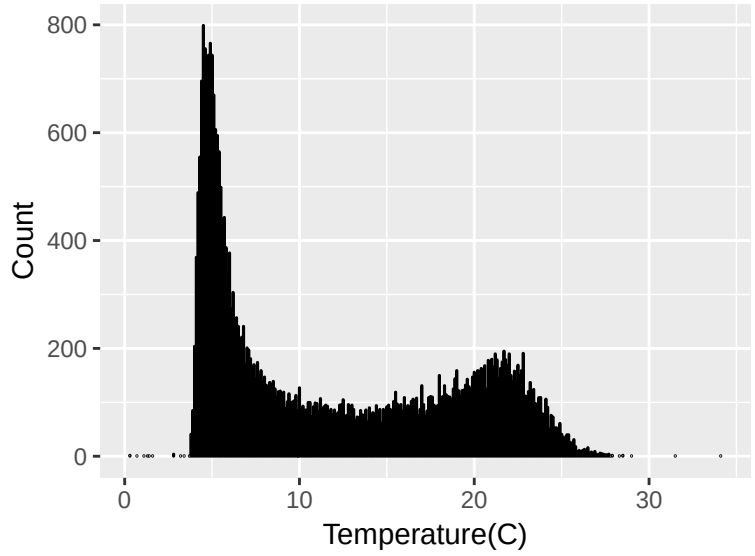
4) Explore your data graphically

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

Bar chart of temperature counts for each lake

```
# 1
library(ggplot2)
Barplot<-ggplot(LakesData, aes(x = temperature_C)) +
  geom_bar(color="black")
Barplot + scale_x_continuous(name="Temperature(C)") +
  scale_y_continuous(name="Count")
```

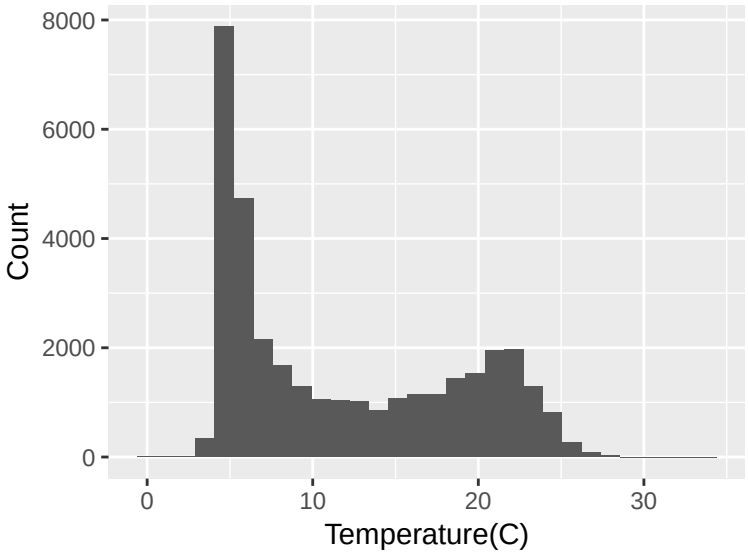


2

Histogram of count distributions of temperature (all temp measurements together)

```
library(ggplot2)
Histogram<-ggplot(LakesData) +
  geom_histogram(aes(x = temperature_C))
Histogram + scale_x_continuous(name="Temperature(C)") +
  scale_y_continuous(name="Count")
```

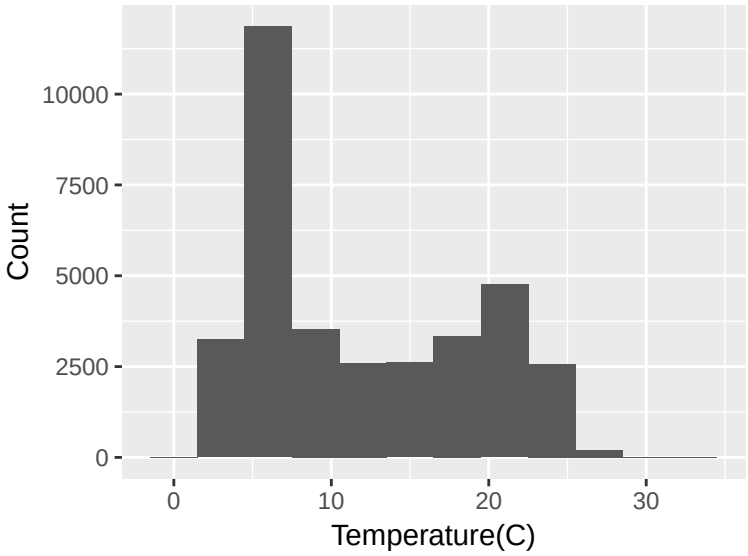
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



3

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

```
HistogramBins<-ggplot(LakesData) +
  geom_histogram(aes(x = temperature_C), binwidth = 3)
HistogramBins +scale_x_continuous(name="Temperature(C)") +
  scale_y_continuous(name="Count")
```



4

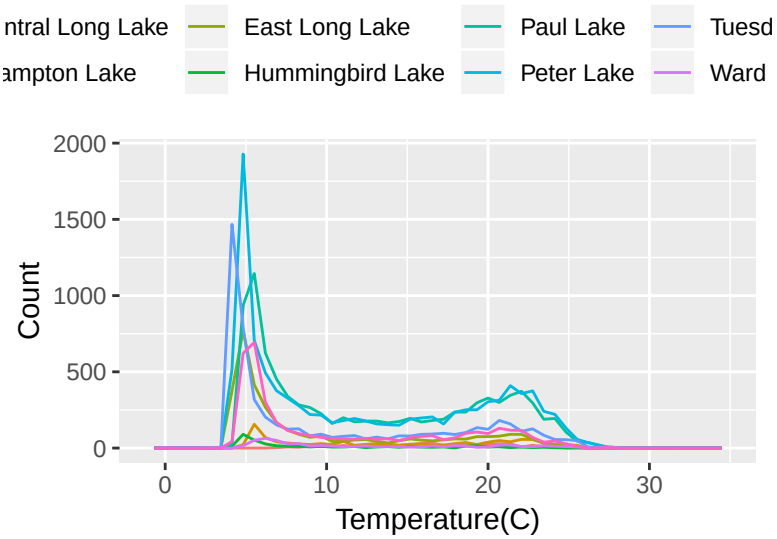
Frequency polygon of temperature for each lake. Choose different colors for each lake.

```
FrequencyPolygon<-ggplot(LakesData) +
  geom_freqpoly(aes(x = temperature_C, color = lakename), bins = 50) +
  scale_x_continuous(limits = c(0, 31)) +
  theme(legend.position = "top")

FrequencyPolygon + scale_x_continuous(name="Temperature(C)") +
  scale_y_continuous(name="Count")
```

```
## Scale for 'x' is already present. Adding another scale for 'x', which
## will replace the existing scale.
```

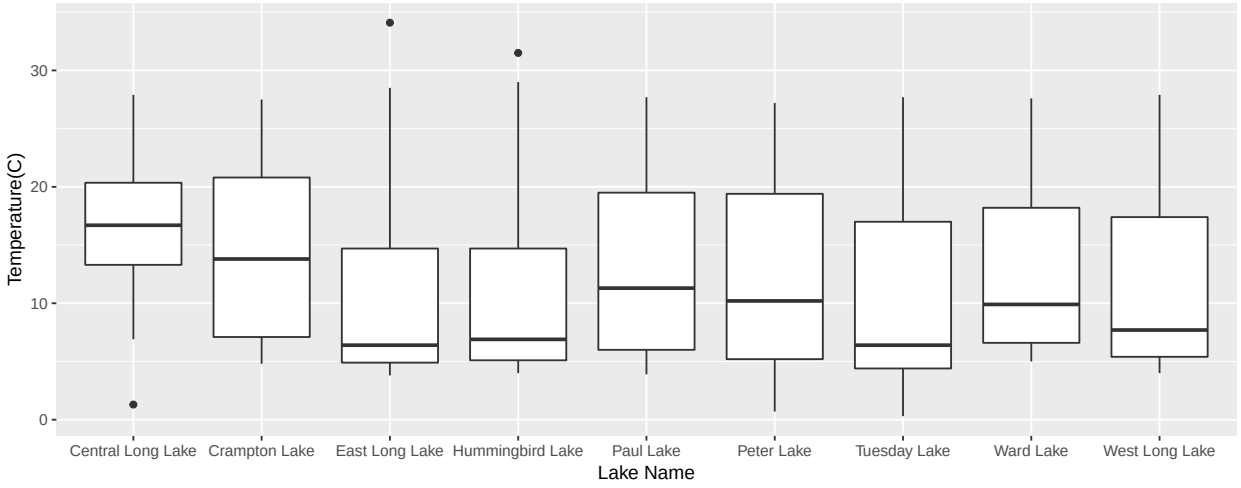
```
## Warning: Removed 3858 rows containing non-finite values (stat_bin).
```



5

Boxplot of temperature for each lake

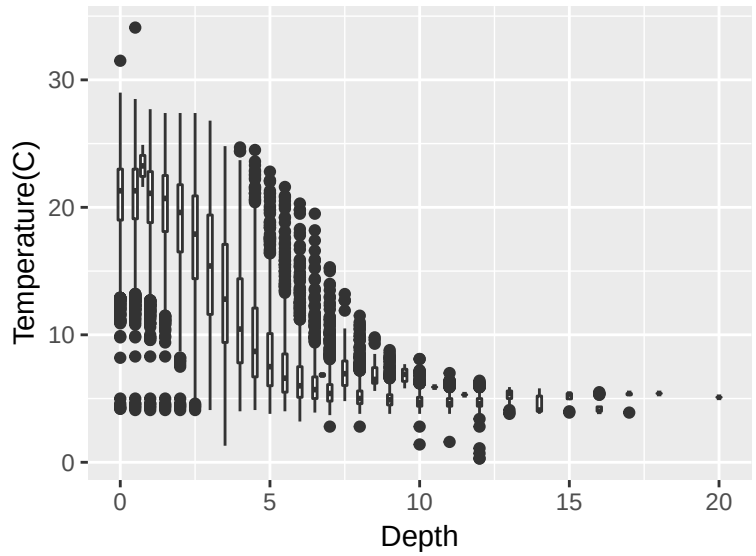
```
ggplot(LakesData) +
  geom_boxplot(aes(x=lakename, y =temperature_C)) +xlab('Lake Name') +ylab('Temperature(C)')
```



6

Boxplot of temperature based on depth, with depth divided into 0.25 m increments

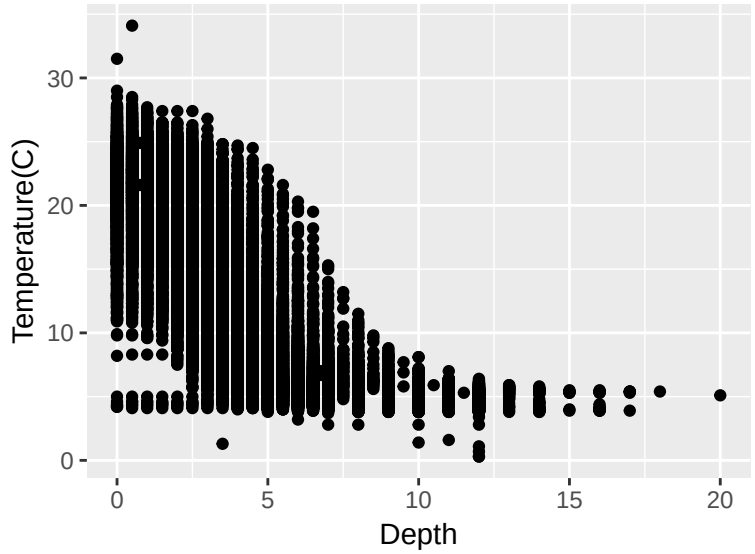
```
ggplot(LakesData) +
  geom_boxplot(aes(x =depth, y =temperature_C, group = cut_width(depth, 0.25))) +xlab('Depth') +ylab('T
```



7

Scatterplot of temperature by depth

```
ggplot(LakesData) +
  geom_point(aes(x =depth, y = temperature_C)) + xlab('Depth') + ylab('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: According to the water samples collected, as the depth of the water samples increases, the temperature of the water samples decreases. This is visually depicted in the two boxplots and the scatterplot I created, and makes sense scientifically because sunlight will reach the top of the lake more easily. Furthermore, it was interesting to note that the majority of the water samples were collected at temperatures below 10 degrees Celsius, based on the bar chart, histograms, and frequency polygon graphs. This corresponds with the ReadMe file which states that the physical and chemical variables were measured at one central station near the deepest point of each lake.

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

> ANSWER 1: What was the research question for this project? It would assist us in determining whether the NA’s should indeed be kept in the dataset or removed. Furthermore, what were the most common reasons for the NA’s being recorded for the different variables? ANSWER 2: What other variables would affect the relationship between depth of the lake and the temperature of the lake, if any? ANSWER 3: Is there a relationship between the temperature and the dissolved oxygen of a water sample? If so, is it a positive or negative relationship?