# Assignment 3: Data Exploration

*Tristen Townsend*

## 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/Tristen/OneDrive - Duke University/Spring 2019/Data Analytics/Environmental_Data_Analytic
```

```
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

## Warning: package 'tibble' was built under R version 3.5.2

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

```
NTL.data <- read.csv("/Users/Tristen/OneDrive - Duke University/Spring 2019/Data Analytics/Environmenta
```

## 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) We are told where/when our data was accessed (North Temperate Lakes Long Term Ecological Research website on 2018-12-06),
>
> 2) how it was downloaded (important for reproducibility-
>
> - Cascade (NTL Categories)
> - Cascade Project at North Temperate Lakes LTER Core Data Carbon 1984 - 2016 AND
> - Cascade Project at North Temperate Lakes LTER Core Data Nutrients 1991 - 2016 AND
> - Cascade Project at North Temperate Lakes LTER Core Data Physical and Chemical Limnology 1984 - 2016
> - On each of the three pages, Download All Data (csv) was chosen.),
>
> 3) and the units our carbon and nutrient data will be in and how the measurements were gathered/read in the field/lab.

## 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.data)
```

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

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

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

```
# 3
head(NTL.data, n=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.data$lakename)
```

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

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

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

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

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

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

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

```
# 5
summary(NTL.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.data$depth)
```

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

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

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

```
head(NTL.data$sampledate, n = 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, because NAs seemed to be scattered across the entire dataset and we wouldn't want to remove any rows with just one NA and lose other information. This indicates that NAs are not likely attributable to measuring errors or the inability to measure for a long period of time. Instead, we can leave the NAs in and if we think NAs might impact a particular analysis, we can tell R to ignore the NAs for that particular function.

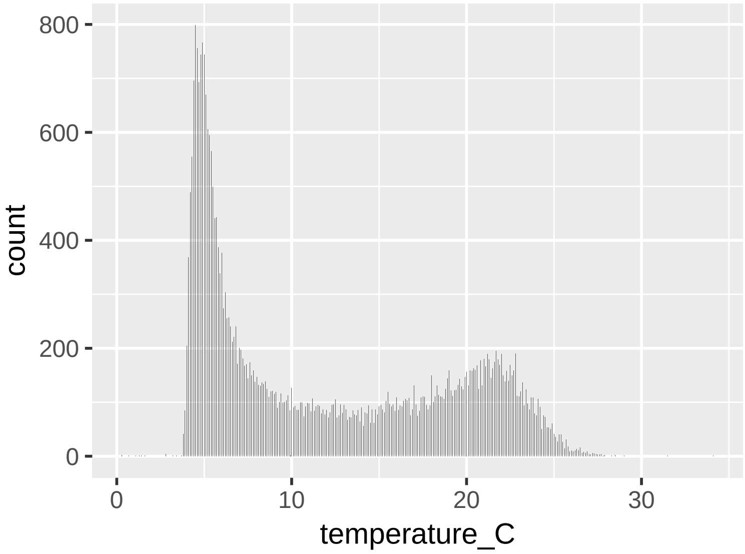
## 4) Explore your data graphically

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 chat of temperature counts for each lake
ggplot(NTL.data, aes(x = temperature_C)) +
  geom_bar()
```

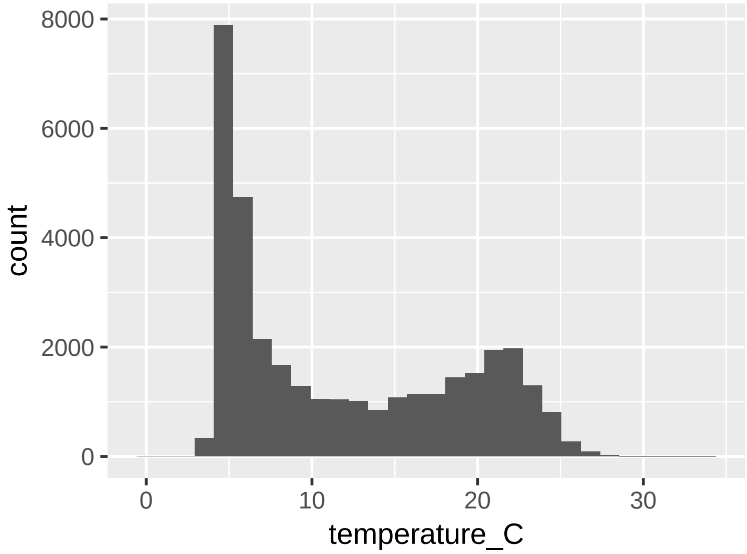
```
## Warning: Removed 3858 rows containing non-finite values (stat_count).
```



```
# 2
ggplot(NTL.data) +
  geom_histogram(aes(x = temperature_C))
```

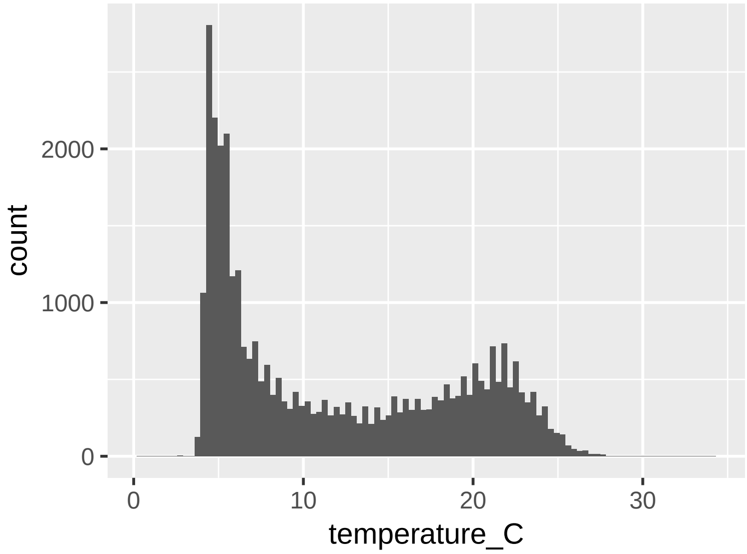
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

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



```
# 3
ggplot(NTL.data) +
  geom_histogram(aes(x = temperature_C), bins = 100)
```

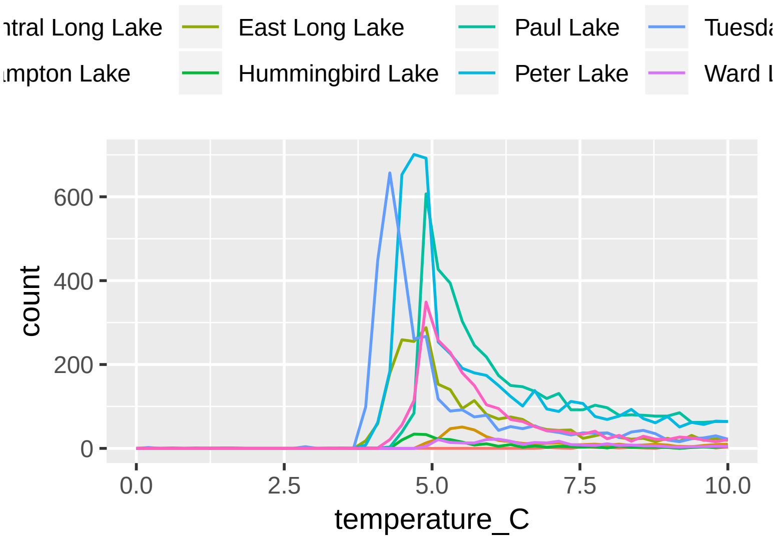
```
## Warning: Removed 3858 rows containing non-finite values (stat_bin).
```



```
# 4
ggplot(NTL.data) +
  geom_freqpoly(aes(x = temperature_C, color = lakename), bins = 50) +
  scale_x_continuous(limits = c(0, 10)) +
  theme(legend.position = "top")
```

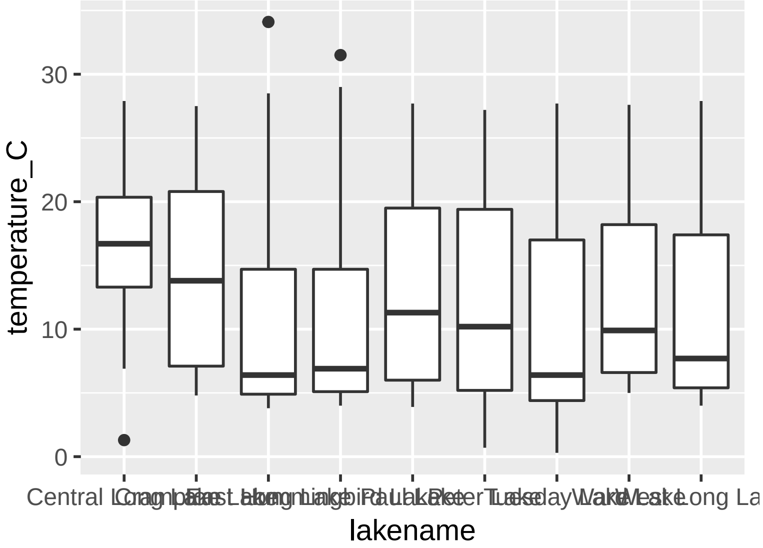
```
## Warning: Removed 20415 rows containing non-finite values (stat_bin).

## Warning: Removed 18 rows containing missing values (geom_path).
```



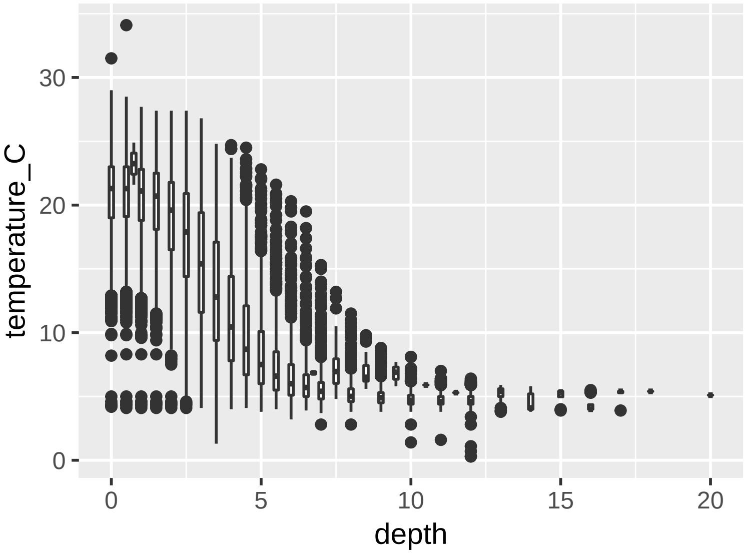
```r
# 5
ggplot(NTL.data) +
  geom_boxplot(aes(x = lakename, y = temperature_C))
```

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



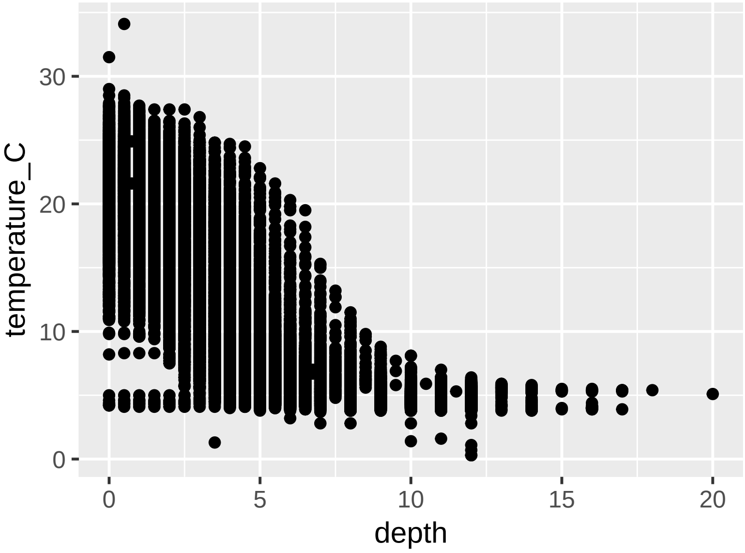
```r
# 6
ggplot(NTL.data) +
  geom_boxplot(aes(x = depth, y = temperature_C, group = cut_width(depth, 0.25)))
```

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



```
# 7
ggplot(NTL.data) +
  geom_point(aes(y = temperature_C, x = depth))
```

```
## 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: The deeper we take measurements, the lower the temperature. Most of the measurements have temperatures around 5 degrees Celsius. East Long Lake and Hummingbird Lake seem to have lower temperatures on average, but both have an outlier that is far higher than any other lake. Also, there is clearly a difference is the temperature range that the different lakes experience.

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

> ANSWER 1: Are East Long Lake and Hummingbird Lake near each other and if so, can we attribute the outliers to some extreme event that may have affected them both?

ANSWER 2: Is the shape of the curve from plot 3 (frequency counts of temperature) attributable to season or a different factor?

ANSWER 3: What is the relationship between temperature and dissolved oxygen? Temperature and irradiance? Are these relationships what we expect, and if not should we collect additional data?