Assignment 3: Data Exploration

Njeri Kara

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.

```
#Setting the working directory
setwd("C:/Users/jerik/OneDrive/Documents/Spring 2019 Semenster/Environmental Data Analytics/EDA_R_Work/
#Confirming that it is the correct working directory
getwd()
```

```
## [1] "C:/Users/jerik/OneDrive/Documents/Spring 2019 Semenster/Environmental Data Analytics/EDA_R_Work/
```

```
#Loading necessary packages
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
## -- Conflicts ------------------------------------------------------------------------------------ tid
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#Uploading the North Temperate Lakes long term monitoring dataset
North.Temp.Lakes.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:
>
> The naming conventions and files format section gives details of the information that can be derived from the data file name. The data file is from the database NTL-LTER, the data described is of lakes, details include chemisty and physics, it is the raw data and it is in csv format.
>
> The data was accessed, 2018-12-06
>
> The data was assembled by Kateri Salk from the North Temperate Lakes Long Term Ecological Research website and Kateri Salk can be contacted at kateri.salk@duke.edu for additional information and support.

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 - Dimensions of the dataset
dim(North.Temp.Lakes.data) #shows the number of rows and columns in the dataset
```

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

```
# 2 - Class of the dataset
class(North.Temp.Lakes.data) #type of dataset - data.frame
```

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

```
# 3 - First eight rows of the dataset
head(North.Temp.Lakes.data,8) #shows first eight rows of the dataset
```

```
##   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(North.Temp.Lakes.data$lakename) #class of the variable lakename - factor
```

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

```
class(North.Temp.Lakes.data$sampledate) #class of the variable sampledate - factor
```

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

```
class(North.Temp.Lakes.data$depth) #class of the variable depth - numeric
```

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

```
class(North.Temp.Lakes.data$temperature_C) #class of the variable temperature_C - numeric
```

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

```
# 5
summary(North.Temp.Lakes.data$lakename) #summary of the variable 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(North.Temp.Lakes.data$depth) #summary of the variable depth
```

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

```
summary(North.Temp.Lakes.data$temperature_C) #summary of the variable temerature_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.

```
#Checked the North.Temp.Lakes.data dataset to confirm the format of the factor variable sample date
#is mm/dd/yy

#Changing the sampledate factor variable to a date variable with the format mm/dd/yy
North.Temp.Lakes.data$sampledate <- as.Date(North.Temp.Lakes.data$sampledate, format = "%m/%d/%y")

#Confirming that the sampledate variable is a date
class(North.Temp.Lakes.data$sampledate) #new class - Date
```

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

Question: Do you want to remove NAs from this dataset? Why or why not?

```
#summary of all the variables in the dataset to see how many NAs are in each variable has
summary(North.Temp.Lakes.data)
```

```
##      lakeid               lakename         year4          daynum
##  R      :11288   Peter Lake    :11288   Min.   :1984   Min.   : 55.0
##  L      :10325   Paul Lake     :10325   1st Qu.:1991   1st Qu.:166.0
##  T      : 6107   Tuesday Lake  : 6107   Median :1997   Median :194.0
##  W      : 4188   West Long Lake: 4188   Mean   :1999   Mean   :194.3
```

```
##  E      : 3905   East Long Lake: 3905   3rd Qu.:2006   3rd Qu.:222.0
##  M      : 1234   Crampton Lake : 1234   Max.   :2016   Max.   :307.0
##  (Other): 1567   (Other)       : 1567
##    sampledate              depth       temperature_C   dissolvedOxygen
##  Min.   :1984-05-27   Min.   : 0.00   Min.   : 0.30   Min.   :  0.00
##  1st Qu.:1991-08-08   1st Qu.: 1.50   1st Qu.: 5.30   1st Qu.:  0.30
##  Median :1997-07-28   Median : 4.00   Median : 9.30   Median :  5.60
##  Mean   :1999-02-05   Mean   : 4.39   Mean   :11.81   Mean   :  4.97
##  3rd Qu.:2006-06-06   3rd Qu.: 6.50   3rd Qu.:18.70   3rd Qu.:  8.40
##  Max.   :2016-08-17   Max.   :20.00   Max.   :34.10   Max.   :802.00
##                                       NA's   :3858    NA's   :4039
##  irradianceWater     irradianceDeck
##  Min.   :   -0.337   Min.   :   1.5
##  1st Qu.:   14.000   1st Qu.: 353.0
##  Median :   65.000   Median : 747.0
##  Mean   :  210.242   Mean   : 720.5
##  3rd Qu.:  265.000   3rd Qu.:1042.0
##  Max.   :24108.000   Max.   :8532.0
##  NA's   :14287       NA's   :15419
##                               comments
##  DO Probe bad - Doesn't go to zero:  206
##  DO taken with Jones Lab Meter    :  162
##  NA's                             :38246
##
##
##
##
```

```
#Removing rows in the dataset with an NA in the temperature_C variable
North.Temp.Lakes.data.no.temp.NAs <- North.Temp.Lakes.data[!is.na(North.Temp.Lakes.data$temperature_C),
#confirming all the rows with an NA in the temperature_C variable have been removed
summary(North.Temp.Lakes.data.no.temp.NAs$temperature_C)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.30    5.30    9.30   11.81   18.70   34.10
```

ANSWER:

I do not want to remove all the NAs in the dataset. This is because variables such as comments, irradianceWater and irradianceDeck have a large proporation of NAs compared to the total number of observations. If they are to be removed, the dataset rows would significantly reduce probably impacting data analysis outcomes.

I do however want to remove the rows with NAs in the temperature_C variable. These rows are just about 10% of the total observations and since temperature_C is used in all the subsequent plots of question 4, it may be beneficial for the temperature_C variable not to have any missing values.

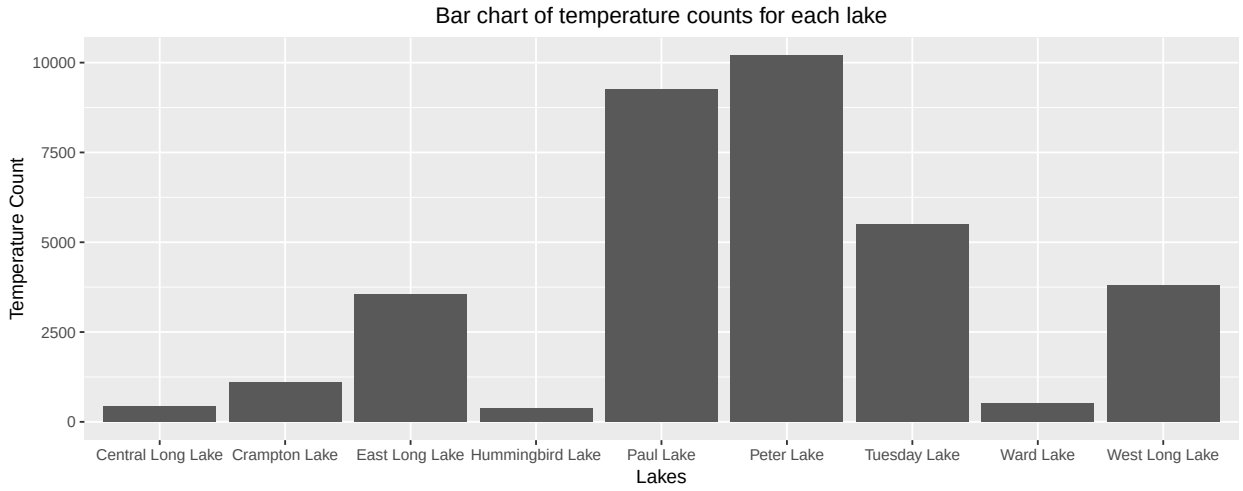
4) Explore your data graphically

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.

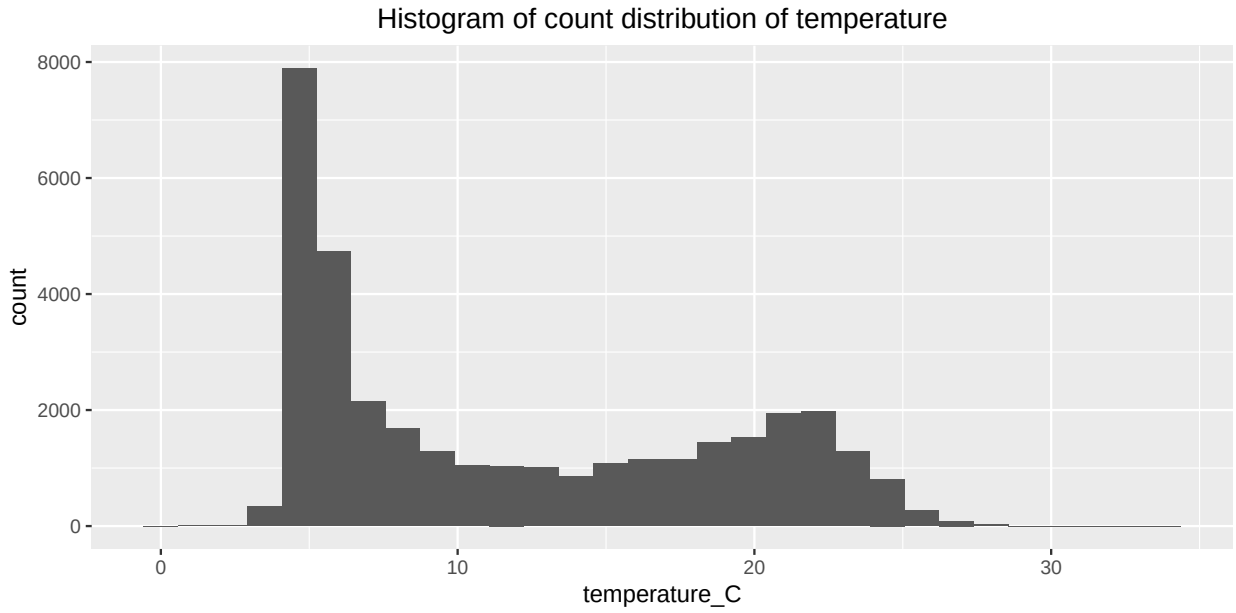
5. Boxplot of temperature for each lake
6. Boxplot of temperature based on depth, with depth divided into 0.25 m increments
7. Scatterplot of temperature by depth

```
# 1. Bar chart of temperature counts for each lake
#The North.Temp.Lakes.data.no.temp.NAs is going to be used because it has only the observations(rows)
#with a temperature value. It can therefore be used to plot the number of temperature readings,
#temperature count, for each lake
ggplot(North.Temp.Lakes.data.no.temp.NAs, aes(x=lakename)) + geom_bar() +
  ggtitle("Bar chart of temperature counts for each lake") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Lakes") + ylab("Temperature Count") #Bar chart with title and labeled x and y axis
```

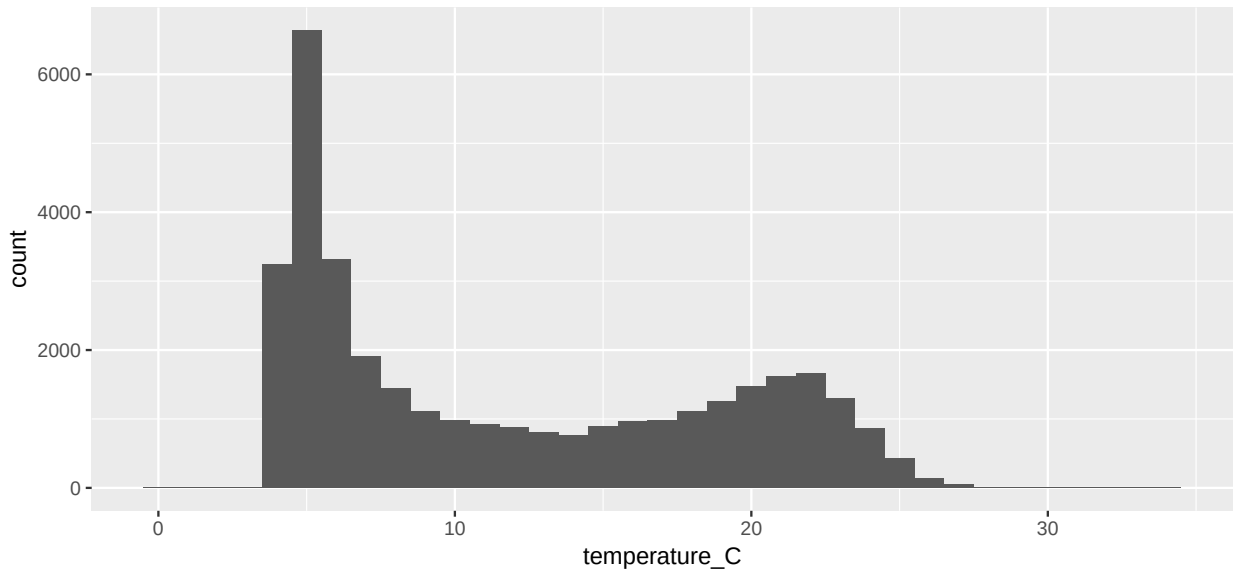



```
# 2. Histogram of count distributions of temperature
ggplot(North.Temp.Lakes.data.no.temp.NAs) +
  geom_histogram(aes(x = temperature_C)) +
  ggtitle("Histogram of count distribution of temperature") +
  theme(plot.title = element_text(hjust = 0.5)) #histogram with a title
```

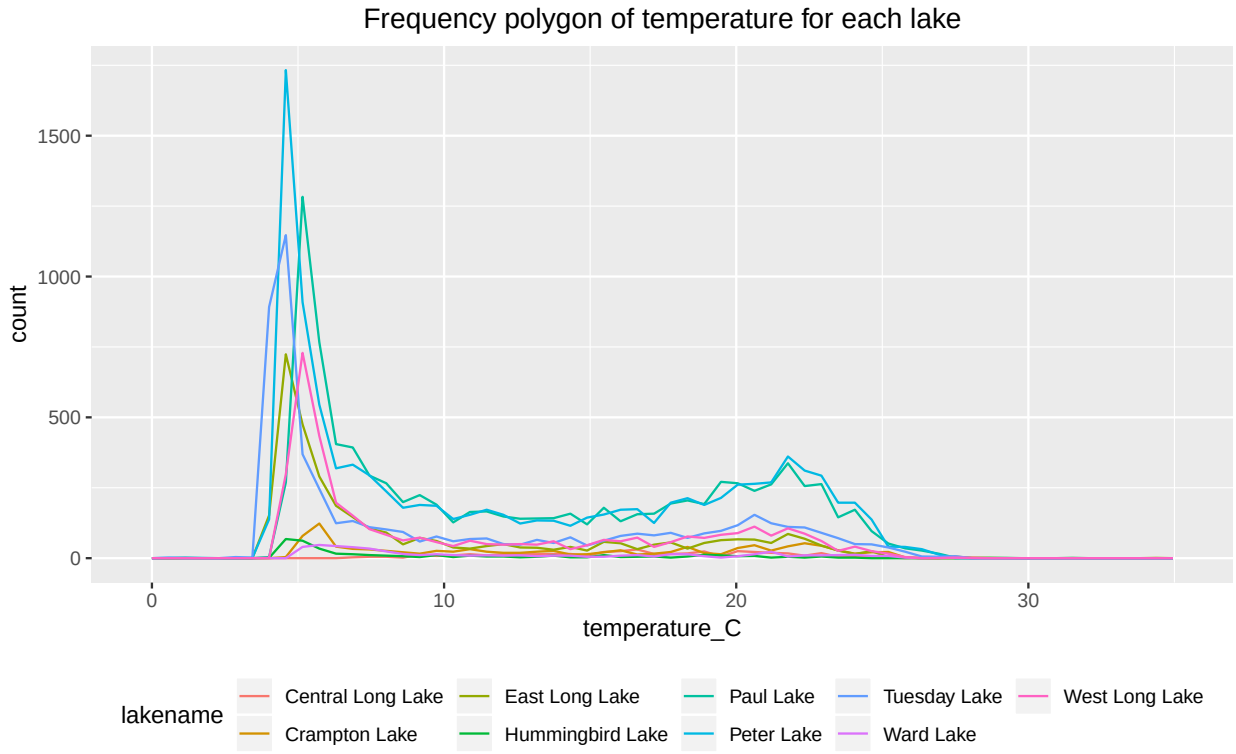
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# 3. Histogram from 2 with a different number of bins
ggplot(North.Temp.Lakes.data.no.temp.NAs, aes(x = temperature_C)) +
  geom_histogram(binwidth = 1) +
  ggtitle("Histogram of count distributions of temperature with binwidths of 1") +
  theme(plot.title = element_text(hjust = 0.5)) #histogram with a binwidth of 1 and a title
```

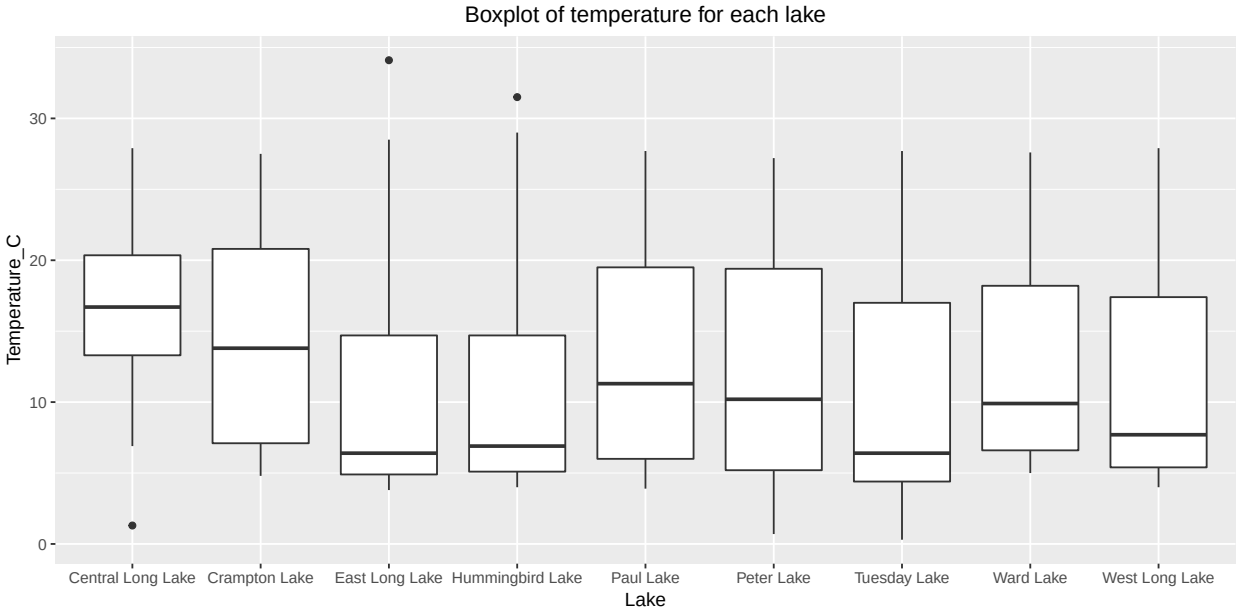



```
# 4. Frequency polygon of temperature for each lake with different colours
ggplot(North.Temp.Lakes.data.no.temp.NAs) +
  geom_freqpoly(aes(x = temperature_C, color = lakename), bins = 60) +
  theme(legend.position = "bottom") + ggtitle("Frequency polygon of temperature for each lake") +
  theme(plot.title = element_text(hjust = 0.5)) #colour of each line is based on the lakename
```

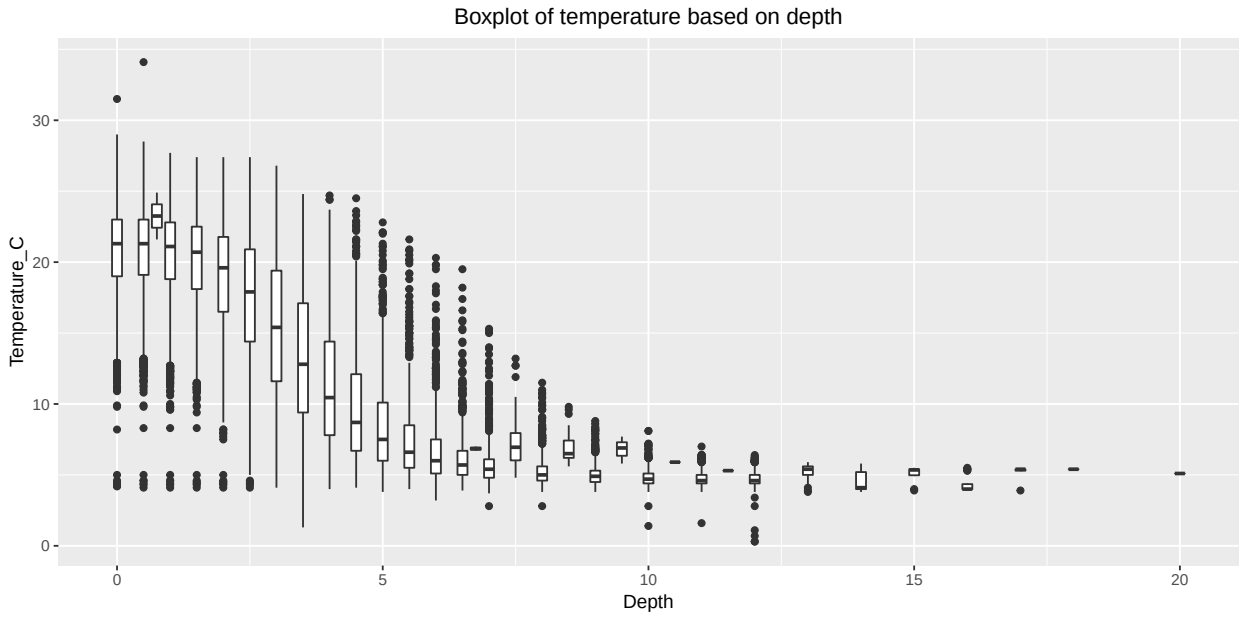



```
#including a legend of the lake colours and a plot title
```

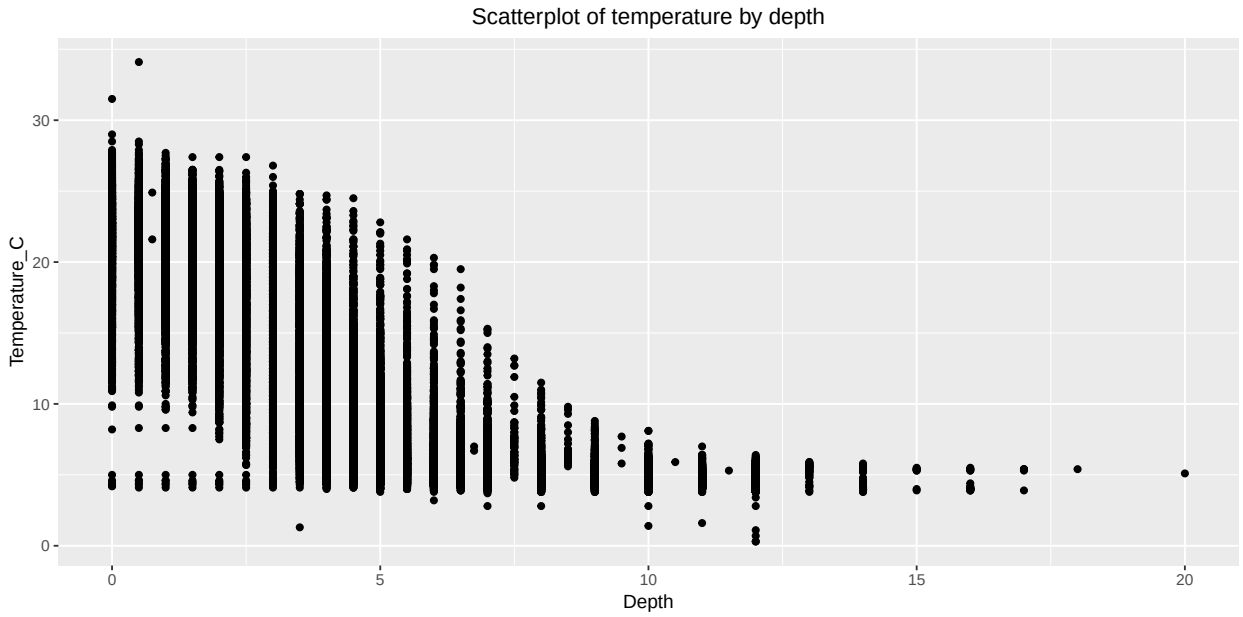
```
# 5. Boxplot of temperature for each lake
ggplot(North.Temp.Lakes.data.no.temp.NAs) +
  geom_boxplot(aes(x = lakename, y = temperature_C))+
  ggtitle("Boxplot of temperature for each lake") + theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Lake") + ylab("Temperature_C") #including a title and axis labels
```



```
# 6. Boxplot of temperature based on depth, with depth divided into 0.25 m increments
ggplot(North.Temp.Lakes.data.no.temp.NAs) +
  geom_boxplot(aes(x = depth, y = temperature_C, group = cut_width(depth, 0.25))) +
  ggtitle("Boxplot of temperature based on depth") +
  theme(plot.title = element_text(hjust = 0.5)) + xlab("Depth") +
  ylab("Temperature_C") #uncluding a plot title and axis labels
```




```
# 7. Scatterplot of temperature by depth
ggplot(North.Temp.Lakes.data.no.temp.NAs) +
  geom_point(aes(x = depth, y = temperature_C)) + ggtitle("Scatterplot of temperature by depth") +
  theme(plot.title = element_text(hjust = 0.5)) + 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:

From the summary of the main dataset I found that only the last 5 variables - temperature_C,dissolvedOxygen irradianceWater,irradianceDeck and comments - have missing values and temperature_C was missing the least number of variables (3858). I therfore decided to carry out by subsequent analysis of temperature with a subset of the dataframe that only had observations with temperature_c values.

The histogram and frequency polygons revealed that the temerature data has a positively skewed distribution. This is consitent with the temperature's summary statistics because its median is less than its mean.

The Bar chart and frequency polygon of temperature counts dissaggregate the temperature count data by lakes clearly showing that Peter Lake has the highest number of temperature observations followed by Paul lake and that the most common temperature reading is approximately 5.

The box plot provides a visual break down of the summary statistics of temperature by lake. It shows for example that the max temperature reading of 34 was taken at the East Long Lake. It also shows that the median value of the temperature readings is mainly determinded by readings from Peter and Paul lakes

The scatterplot and boxplot of temperature based on depth reveal that the range and number of temperature readings reduces with increasing depth.

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

ANSWER 1: The number of temperature readings taken every year from 1984 to 2016 and the historical trend of this temperature data collection by year.

ANSWER 2: The change in depth observations as years progressed from 1984 to 2016

ANSWER 3: The relationship between obseration depth and the lake the observation is taken from.