Assignment 7: GLMs (Linear Regressios, ANOVA, & t-tests)

Fiona Kelley

Fall 2023

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on generalized linear models.

Directions

1. Rename this file `<FirstLast>_A07_GLMs.Rmd` (replacing `<FirstLast>` with your first and last name).
2. Change "Student Name" on line 3 (above) with your name.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.

Set up your session

1. Set up your session. Check your working directory. Load the tidyverse, agricolae and other needed packages. Import the *raw* NTL-LTER raw data file for chemistry/physics (`NTL-LTER_Lake_ChemistryPhysics_Raw.csv`). Set date columns to date objects.

2. Build a ggplot theme and set it as your default theme.

```
#1 Loading the necessary packages needed for this assignment.

library(tidyverse)
library(agricolae)
library(ggplot2)
library(lubridate)
library(here)
library(dplyr)

#Checking the working directory.
getwd()
```

```
## [1] "/home/guest/R/EDA-Spring2023"
```

```
#Importing the data set for this assignment.
Lake_Chem_Physics <- read.csv("./Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv",
  stringsAsFactors = TRUE)
```

```
#Setting the year column to a date object.
Lake_Chem_Physics$sampledate <- as.Date(Lake_Chem_Physics$sampledate, format = "%m/%d/%Y")

#2 Creating a ggplot theme and setting it as the default.

MyTheme <- theme_classic() + theme(
        #Sets text on the axes as black.
        axis.text = element_text(color = 'black'),
        #Positions the legend to the right of plots.
        legend.position = ("right"),
        #Outlines and fills the legend pink.
        legend.background = element_rect(color = 'pink', fill = 'pink'),
        #Outlines and fills the background grey.
        plot.background = element_rect(color = 'grey', fill = 'grey'),
        #Sets grid lines for the plot as grey.
        panel.grid.minor = element_line(color = "grey"),
        panel.grid.major = element_line(color = "grey")
)

#Setting MyTheme as the default theme.
theme_set(MyTheme)
```

Simple regression

Our first research question is: Does mean lake temperature recorded during July change with depth across all lakes?

3. State the null and alternative hypotheses for this question: > Answer: H0: Mean lake temperature does not change with depth across all lakes in July. Ha: Mean lake temperature does change with depth across all lakes in July.
4. Wrangle your NTL-LTER dataset with a pipe function so that the records meet the following criteria:

- Only dates in July.
- Only the columns: `lakename`, `year4`, `daynum`, `depth`, `temperature_C`
- Only complete cases (i.e., remove NAs)

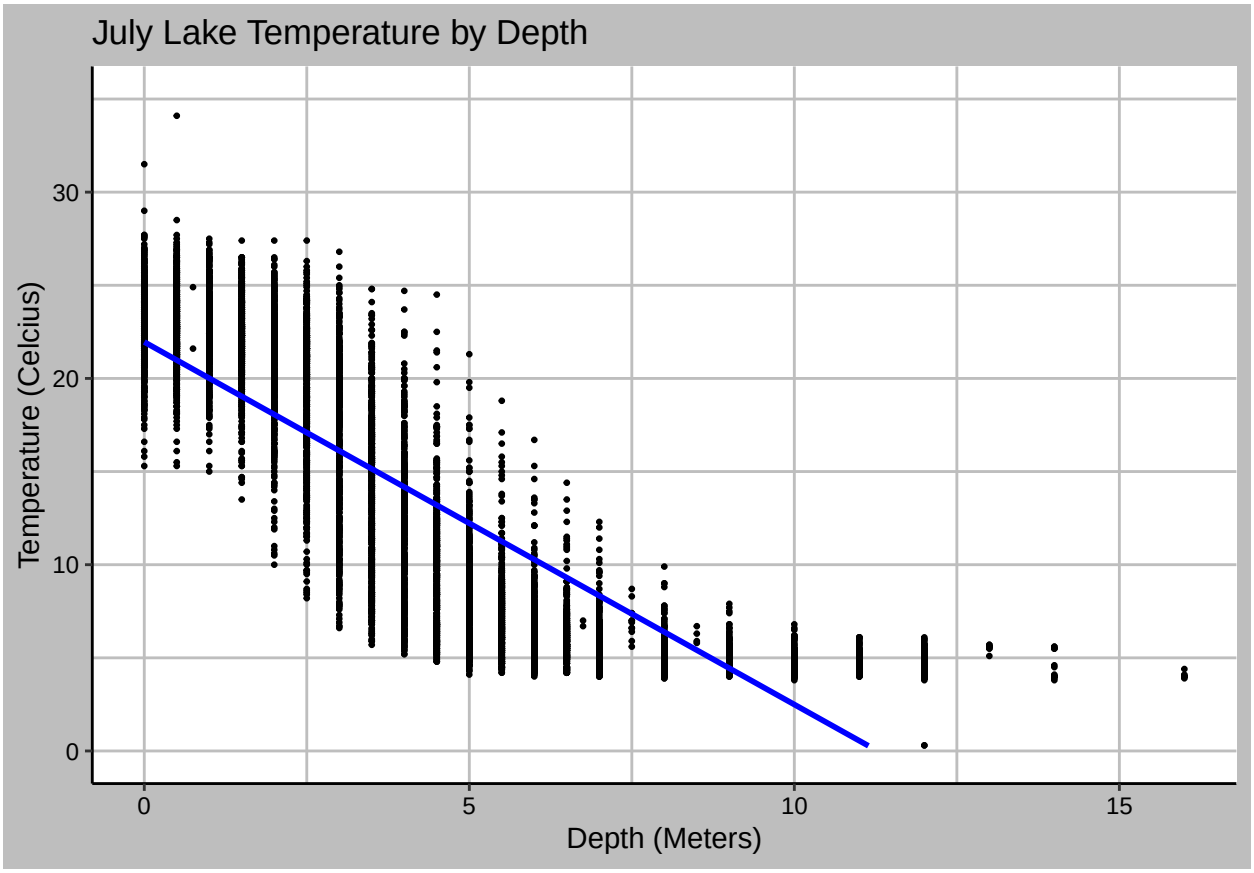
5. Visualize the relationship among the two continuous variables with a scatter plot of temperature by depth. Add a smoothed line showing the linear model, and limit temperature values from 0 to 35 °C. Make this plot look pretty and easy to read.

```
#4 Wrangling the data set to meet the desired criteria.

July_data <- Lake_Chem_Physics %>%
  filter(month(sampledate) == 7) %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  drop_na()


#5 Creating a scatter plot for the wrangled data.

Lake_scatterplot <- ggplot(July_data, aes(x = depth, y = temperature_C)) +
```

```
  geom_point(size = 0.5)+ ylim(c(0, 35)) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(x = "Depth (Meters)",
       y = "Temperature (Celcius)",
       title = "July Lake Temperature by Depth")

Lake_scatterplot
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 24 rows containing missing values (`geom_smooth()`).
```



6. Interpret the figure. What does it suggest with regards to the response of temperature to depth? Do the distribution of points suggest about anything about the linearity of this trend?

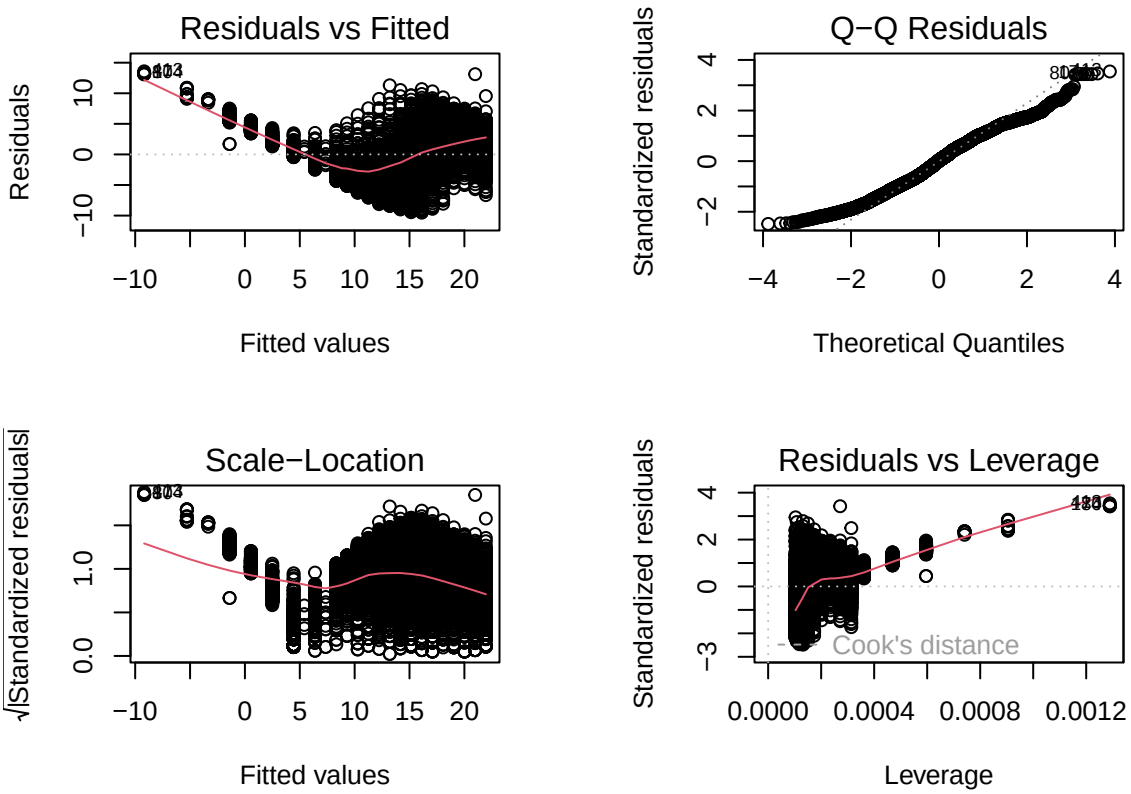
   Answer: The relationship present between temperature and depth is clearly displayed by the trend line on the scatter plot as a negative correlation. As depth increases, temperature decreases, revealing an inverse realtionship between the two variables. The distribution of points do not follow a linear pattern, suggesting a singular linear regression for all the lakes combined may not be the most appropriate.

7. Perform a linear regression to test the relationship and display the results

```r
#7 Running a linear regression for depth by temperature.

lake_regression <- lm(July_data$temperature_C ~ July_data$depth)

#Summarizing the regression.
summary(lake_regression)
```

```
##
## Call:
## lm(formula = July_data$temperature_C ~ July_data$depth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5173 -3.0192  0.0633  2.9365 13.5834
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     21.95597    0.06792   323.3   <2e-16 ***
## July_data$depth -1.94621    0.01174  -165.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.835 on 9726 degrees of freedom
## Multiple R-squared:  0.7387, Adjusted R-squared:  0.7387
## F-statistic: 2.75e+04 on 1 and 9726 DF,  p-value: < 2.2e-16
```

```r
#Displaying the data.
par(mfrow = c(2,2), mar=c(4,4,4,4))
plot(lake_regression)
```




```
par(mfrow = c(1,1))
```

8. Interpret your model results in words. Include how much of the variability in temperature is explained by changes in depth, the degrees of freedom on which this finding is based, and the statistical significance of the result. Also mention how much temperature is predicted to change for every 1m change in depth.

   Answer: The calculated p-value (<2e-16) was less than the significance level (0.05), indicating the results are statisically significant. As a result, we reject the null hypothesis. The high value of the Multiple R-squared (0.7387) reveals a strong correlation between depth and temperature. In other words, changes in lake depth explains about 73.87 percent of the variability in temperature. The results of this linear regression were based on 9,726 degrees of freedom. For every 1 meter change in depth, temperature decreases by about 1.95 degrees Celcius.

Multiple regression

Let's tackle a similar question from a different approach. Here, we want to explore what might the best set of predictors for lake temperature in July across the monitoring period at the North Temperate Lakes LTER.

9. Run an AIC to determine what set of explanatory variables (year4, daynum, depth) is best suited to predict temperature.

10. Run a multiple regression on the recommended set of variables.

```
#9 Running an AIC to determine which variable is best suited to predict temperature.

temp_AIC <- lm(data = July_data, temperature_C ~ year4 + daynum + depth)
#Choosing a model by AIC in a Stepwise Algorithm.
step(temp_AIC)
```

```
## Start:  AIC=26065.53
## temperature_C ~ year4 + daynum + depth
##
##          Df Sum of Sq    RSS   AIC
## <none>                141687 26066
## - year4   1       101 141788 26070
## - daynum  1      1237 142924 26148
## - depth   1    404475 546161 39189

##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = July_data)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##    -8.57556      0.01134      0.03978     -1.94644
```

```
#10 Running a multiple regression.
temp_regression <- lm(data = July_data, temperature_C ~ year4 + daynum + depth)
# Summarizing the multiple regression.
summary(temp_regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = July_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6536 -3.0000  0.0902  2.9658 13.6123
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -8.575564   8.630715   -0.994  0.32044
## year4        0.011345   0.004299    2.639  0.00833 **
## daynum       0.039780   0.004317    9.215  < 2e-16 ***
## depth       -1.946437   0.011683 -166.611  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.817 on 9724 degrees of freedom
## Multiple R-squared:  0.7412, Adjusted R-squared:  0.7411
## F-statistic:  9283 on 3 and 9724 DF,  p-value: < 2.2e-16
```

11. What is the final set of explanatory variables that the AIC method suggests we use to predict temperature in our multiple regression? How much of the observed variance does this model explain? Is this an improvement over the model using only depth as the explanatory variable?

Answer: The AIC method suggests we should use all three variables to predict temperature in our multiple regression. The Stepwise algorithm revealed the lowest AIC value for "none", meaning keeping all three explanatory variables will provide the best prediction of temperature. After running the multiple regression, the R-squared value (0.7412) slightly increased from the depth linear regression (0.7387), indicating a stronger relationship between temperature and the three variables. The R-Squared value of 0.7412 is the proportion of the observed variance explained by the model.

Analysis of Variance

12. Now we want to see whether the different lakes have, on average, different temperatures in the month of July. Run an ANOVA test to complete this analysis. (No need to test assumptions of normality or similar variances.) Create two sets of models: one expressed as an ANOVA models and another expressed as a linear model (as done in our lessons).

```
#12 Creating an ANOVA model and linear model.

# ANOVA model
Lake_anova <- aov(data = July_data, temperature_C ~ lakename)
summary(Lake_anova)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## lakename       8  21642  2705.2      50 <2e-16 ***
## Residuals   9719 525813    54.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Linear model
Lake_anova2 <- lm(data = July_data, temperature_C ~ lakename)
summary(Lake_anova2)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = July_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.769  -6.614  -2.679   7.684  23.832
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 17.6664     0.6501  27.174  < 2e-16 ***
## lakenameCrampton Lake       -2.3145     0.7699  -3.006 0.002653 **
## lakenameEast Long Lake      -7.3987     0.6918 -10.695  < 2e-16 ***
## lakenameHummingbird Lake    -6.8931     0.9429  -7.311 2.87e-13 ***
## lakenamePaul Lake           -3.8522     0.6656  -5.788 7.36e-09 ***
## lakenamePeter Lake          -4.3501     0.6645  -6.547 6.17e-11 ***
## lakenameTuesday Lake        -6.5972     0.6769  -9.746  < 2e-16 ***
## lakenameWard Lake           -3.2078     0.9429  -3.402 0.000672 ***
## lakenameWest Long Lake      -6.0878     0.6895  -8.829  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.355 on 9719 degrees of freedom
## Multiple R-squared:  0.03953,    Adjusted R-squared:  0.03874
## F-statistic:    50 on 8 and 9719 DF,  p-value: < 2.2e-16
```

13. Is there a significant difference in mean temperature among the lakes? Report your findings.

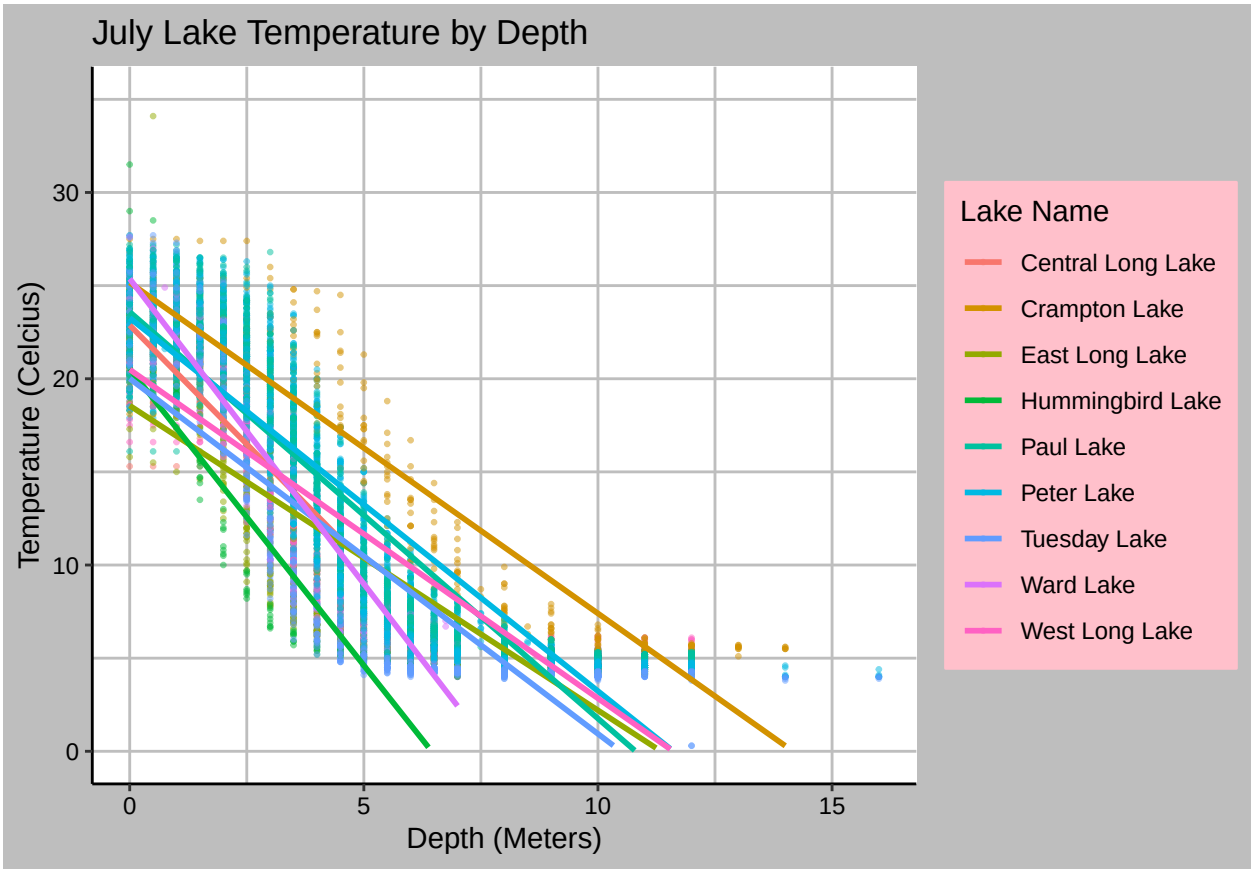
    Answer: The p-value (<2e-16) calculated in the ANOVA model is less than the significance level (0.05), suggesting there is a significant difference in mean temperature among the lakes. Therefore we would reject the null hypothesis.

14. Create a graph that depicts temperature by depth, with a separate color for each lake. Add a geom_smooth (method = "lm", se = FALSE) for each lake. Make your points 50 % transparent. Adjust your y axis limits to go from 0 to 35 degrees. Clean up your graph to make it pretty.

```
#14. Creating a scatter plot meeting the desired criteria.

Lake_scatterplot2 <- ggplot(July_data,
                            aes(x = depth,
                                y = temperature_C,
                                color = lakename)) +
  geom_point(size = 0.5, alpha=0.5) + ylim(c(0, 35)) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Depth (Meters)",
       y = "Temperature (Celcius)",
       title = "July Lake Temperature by Depth",
       color = "Lake Name")

Lake_scatterplot2
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 73 rows containing missing values (`geom_smooth()`).
```



15. Use the Tukey's HSD test to determine which lakes have different means.

```
#15 Using the Tukey's HSD test to examine which lakes have different means.

Lake_tukey <- HSD.test(Lake_anova, "lakename", group = TRUE)
Lake_tukey
```

```
## $statistics
##   MSerror   Df     Mean       CV
##   54.1016 9719 12.72087 57.82135
##
## $parameters
##    test   name.t ntr StudentizedRange alpha
##   Tukey lakename   9         4.387504  0.05
##
## $means
##                   temperature_C      std    r        se Min  Max    Q25   Q50
## Central Long Lake      17.66641 4.196292  128 0.6501298 8.9 26.8 14.400 18.40
## Crampton Lake          15.35189 7.244773  318 0.4124692 5.0 27.5  7.525 16.90
## East Long Lake         10.26767 6.766804  968 0.2364108 4.2 34.1  4.975  6.50
## Hummingbird Lake       10.77328 7.017845  116 0.6829298 4.0 31.5  5.200  7.00
## Paul Lake              13.81426 7.296928 2660 0.1426147 4.7 27.7  6.500 12.40
## Peter Lake             13.31626 7.669758 2872 0.1372501 4.0 27.0  5.600 11.40
## Tuesday Lake           11.06923 7.698687 1524 0.1884137 0.3 27.7  4.400  6.80
## Ward Lake              14.45862 7.409079  116 0.6829298 5.7 27.6  7.200 12.55
```

```
## West Long Lake         11.57865 6.980789 1026 0.2296314 4.0 25.7  5.400  8.00
##                       Q75
## Central Long Lake 21.000
## Crampton Lake     22.300
## East Long Lake    15.925
## Hummingbird Lake  15.625
## Paul Lake         21.400
## Peter Lake        21.500
## Tuesday Lake      19.400
## Ward Lake         23.200
## West Long Lake    18.800
##
## $comparison
## NULL
##
## $groups
##                   temperature_C groups
## Central Long Lake      17.66641      a
## Crampton Lake          15.35189     ab
## Ward Lake              14.45862     bc
## Paul Lake              13.81426      c
## Peter Lake             13.31626      c
## West Long Lake         11.57865      d
## Tuesday Lake           11.06923     de
## Hummingbird Lake       10.77328     de
## East Long Lake         10.26767      e
##
## attr(,"class")
## [1] "group"
```

16.From the findings above, which lakes have the same mean temperature, statistically speaking, as Peter Lake? Does any lake have a mean temperature that is statistically distinct from all the other lakes?

> Answer: Peter Lake shares the same mean temperature, statistically speaking, with Paul Lake and Ward Lake. No lake has a mean temperature that is statistically distinct from all the other lakes.

17. If we were just looking at Peter Lake and Paul Lake. What's another test we might explore to see whether they have distinct mean temperatures?

> Answer: If we were just looking at Peter Lake and Paul Lake, another test we might explore to see whether they have distinct mean temperatures is a two sample t-test because this examines whether two mean values are equivalent.

18. Wrangle the July data to include only records for Crampton Lake and Ward Lake. Run the two-sample T-test on these data to determine whether their July temperature are same or different. What does the test say? Are the mean temperatures for the lakes equal? Does that match you answer for part 16?

```
# Isolating data from Crampton Lake and Ward Lake.
Crampton_Ward_data <- July_data %>%
  filter(lakename == c("Crampton Lake", "Ward Lake"))
```

```r
#Running a two sample t-test for temperature and lake name.
Crampton_Ward_twosample <- t.test(Crampton_Ward_data$temperature_C ~
                                    Crampton_Ward_data$lakename)
Crampton_Ward_twosample
```

```
##
##  Welch Two Sample t-test
##
## data:  Crampton_Ward_data$temperature_C by Crampton_Ward_data$lakename
## t = 0.98673, df = 95.77, p-value = 0.3263
## alternative hypothesis: true difference in means between group Crampton Lake and group Ward Lake is n
## 95 percent confidence interval:
##  -1.130614  3.365610
## sample estimates:
## mean in group Crampton Lake     mean in group Ward Lake
##                    15.37107                    14.25357
```

Answer: The p-value of the two sample t-test performed is greater than the significance level (0.05), indicating a failure to reject the null hypothesis. The mean temperatures for Crampton Lake and Ward Lake are not equal, but the values are rather close. This does match the answer to number 16, stating the two mean temperatures are not statistically significant. The two sample test supports this conclusion because the mean values possess a difference of about 1.