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

Chunyi Xu

Spring 2025

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 necessary packages (tidyverse, lubridate, here)
library (tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library (lubridate)
library (agricolae)
```

```
## Warning: package 'agricolae' was built under R version 4.3.3
```

```r
library (ggplot2)
library(corrplot)
```

```
## Warning: package ’corrplot’ was built under R version 4.3.3

## corrplot 0.95 loaded
```

```r
library (dplyr)
library (here)
```

```
## Warning: package ’here’ was built under R version 4.3.3

## here() starts at /Users/xuchunyi/Desktop/EDA_Spring2025/EDA_Spring2025
```

```r
#Checking working directory is the project folder
here()
```

```
## [1] "/Users/xuchunyi/Desktop/EDA_Spring2025/EDA_Spring2025"
```

```r
#Importing the NTL-LTER raw data file
NTL_LTER <- read.csv(
  file = here("./Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv"),
  stringsAsFactors = TRUE)

#Setting date columns to date objects
NTL_LTER$sampledate <- as.Date(NTL_LTER$sampledate, format = "%m/%d/%y")


#2
#Building my default theme
my.default.theme <- theme_classic(base_size = 15) +
  theme(plot.background = element_rect(color='white',fill='white'),
        plot.title = element_text(color='black', size = 15),
        axis.text = element_text(color = 'black',size = 10),
        panel.grid.minor = element_line(size = 0.5),
        panel.grid.major = element_line(size = 0.5),
        legend.background = element_rect(color='white', fill = 'white'),
        legend.position = "right",
        legend.title = element_text(color='black',size=10),
        legend.text = element_text(size = 10))
```

```
## Warning: The ‘size‘ argument of ‘element_line()‘ is deprecated as of ggplot2 3.4.0.
## i Please use the ‘linewidth‘ argument instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```r
#Setting the theme as the default theme
theme_set(my.default.theme)
```

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: The depth of the lake has no correlation with the lake temparature recorded during July Ha: The depth of the lake has correlation with the lake temparature recorded during 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.

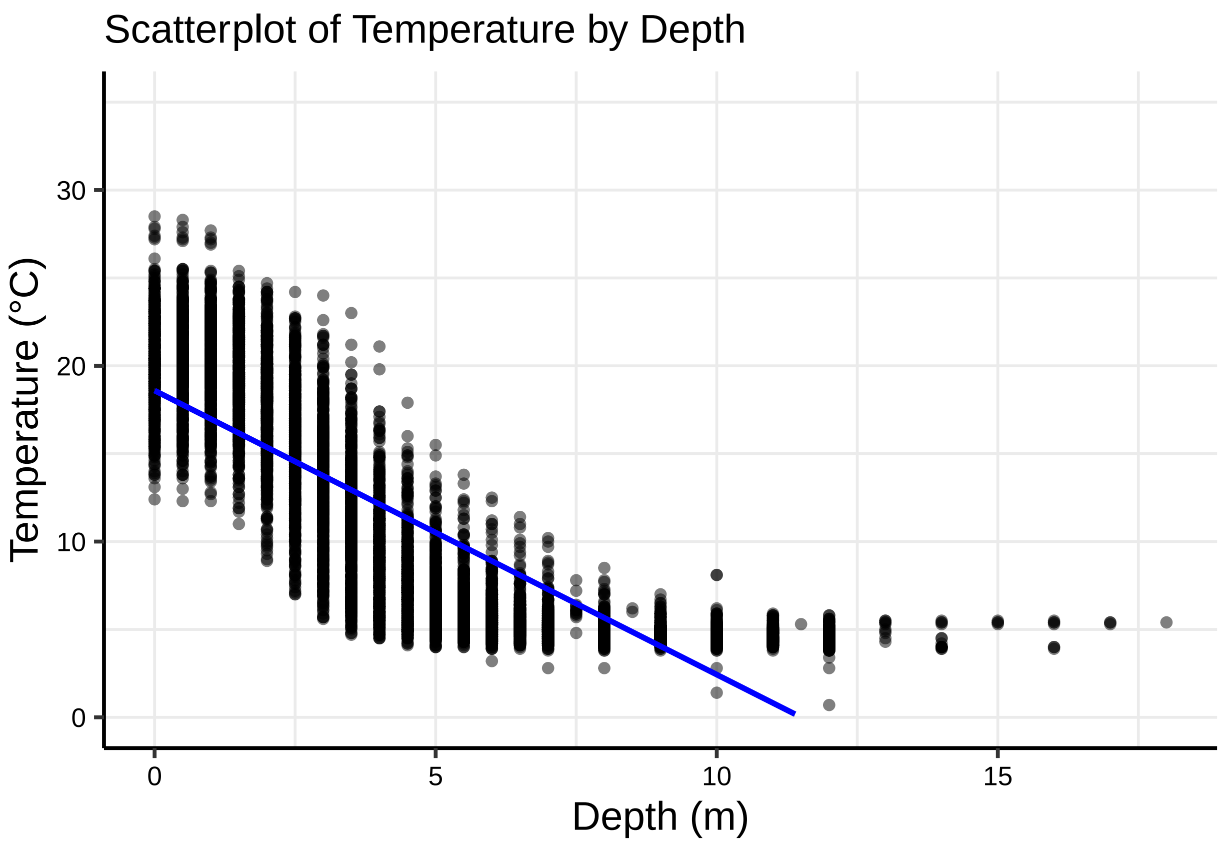
```r
#4
#Wrangling the NTL-LTER dataset to follow the criteria
NTL_LTER_Sub  <- NTL_LTER %>%
  mutate(Month = month(sampledate))%>%
  filter (Month == "6") %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  drop_na()


#5
#Visualizing the relationship among the temperature and depth with a scatter plot
Temp.depth.scatt <-
  ggplot(NTL_LTER_Sub, aes(x = depth, y = temperature_C)) +
  geom_point(alpha = 0.5) +
  geom_smooth (method="lm", color = "blue")+
  labs(x = "Depth (m)", y = "Temperature (°C)") +
  labs(title = "Scatterplot of Temperature by Depth")+
  ylim(0, 35)

print(Temp.depth.scatt)
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```

```
## Warning: Removed 29 rows containing missing values or values outside the scale range
## (‘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: As the depth increases, the lake temperature decreases. The figure suggests temperature is negatively corrleated with depth (negative linearity).

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

```
#7
#Performing a linear regression on temperature and depth
LR.Temp.depth <-
  lm(NTL_LTER_Sub$temperature_C ~
       NTL_LTER_Sub$depth)
summary (LR.Temp.depth)
```

```
##
## Call:
## lm(formula = NTL_LTER_Sub$temperature_C ~ NTL_LTER_Sub$depth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2359 -2.8873 -0.2792  2.6694 15.8990
##
## Coefficients:
```

```
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         18.59256    0.06427   289.3   <2e-16 ***
## NTL_LTER_Sub$depth -1.61620    0.01100  -146.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.582 on 9501 degrees of freedom
## Multiple R-squared:  0.6943, Adjusted R-squared:  0.6942
## F-statistic: 2.158e+04 on 1 and 9501 DF,  p-value: < 2.2e-16
```

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:Since the adjusted R-squared is 0.6942, it suggests 69.42% of the variability in temperature is explained by changes in depth. The degrees of freedom on which this finding is based is 9501. The statistical significance of the result is $<$ 2.2e-16, which is less than 0.05. This p-value suggests that there is significant correlation between temperature and depth. It is predicted 1.62 °C temperature will decrease for every 1m increase in depth.

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 what set of explanatory variables is best suited to predict temperature
Temp.AIC <- lm(data = NTL_LTER_Sub, temperature_C ~ year4 + daynum +
              depth)
step (Temp.AIC)
```

```
## Start:  AIC=23932.45
## temperature_C ~ year4 + daynum + depth
##
##          Df Sum of Sq    RSS   AIC
## <none>                117822 23932
## - year4   1        31 117853 23933
## - daynum  1      4040 121862 24251
## - depth   1    276513 394335 35410
```

```
##
## Call:
```

```
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL_LTER_Sub)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##   18.805558    -0.006318     0.074440    -1.615432
```

```
#Since the smaller the AIC value, the better. The model that includes all three
# explanatory variables (year4, daynum, depth) has the lowest value of AIC,
# which is 23932.
#Therefore, a multiple regression that includes year4, daynum, depth is best suited
#to predict temperature.

#10
#Running a multiple regression on the recommended set of variables
Temp.Mul.Reg <- lm(data = NTL_LTER_Sub, temperature_C ~ year4 + daynum + depth)
summary(Temp.Mul.Reg)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL_LTER_Sub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6228 -2.8170 -0.1937  2.7410 15.7528
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 18.805558   7.973540    2.358   0.0184 *
## year4       -0.006318   0.003970   -1.591   0.1116
## daynum       0.074440   0.004125   18.047   <2e-16 ***
## depth       -1.615432   0.010819 -149.308   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.522 on 9499 degrees of freedom
## Multiple R-squared:  0.7045, Adjusted R-squared:  0.7044
## F-statistic:  7549 on 3 and 9499 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 final set of explanatory variables that the AIC method suggests we use to predict temperature in our multiple regression include year4, daynum, and depth. Since the adjusted R-squared is 0.7044, 70.44% of the observed variance this model can explain. Yes, this is an improvement over the model using only depth as the explanatory variable. The previous model using only depth can only explain 69.42% of the variability in temperature. However, the new multiple regression model can explain more variability in temperature by including additional explanatory variables, which are year4, daynum, and depth.

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).

```r
#12
#H0: The mean of the lake temperature is the same across different lake sites.
#Ha: The mean of the lake temperature is not the same across different lake sites.

#Formatting as an ANOVA model
Temp.anova <- aov(data = NTL_LTER_Sub, temperature_C ~ lakename)
summary(Temp.anova)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)    
## lakename       8  12019  1502.4   36.88 <2e-16 ***
## Residuals   9494 386708    40.7                   
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```r
#Formatting as a linear model
Temp.anova.lm <- lm(data = NTL_LTER_Sub, temperature_C ~ lakename)
summary(Temp.anova.lm)
```

```
## 
## Call:
## lm(formula = temperature_C ~ lakename, data = NTL_LTER_Sub)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -10.405  -5.358  -2.914   5.886  19.302 
## 
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)               15.6992     0.5802  27.058  < 2e-16 ***
## lakenameCrampton Lake     -2.8291     0.6918  -4.090 4.36e-05 ***
## lakenameEast Long Lake    -6.5010     0.6169 -10.538  < 2e-16 ***
## lakenameHummingbird Lake  -6.7827     0.8428  -8.048 9.44e-16 ***
## lakenamePaul Lake         -4.0856     0.5935  -6.884 6.20e-12 ***
## lakenamePeter Lake        -4.5938     0.5926  -7.752 9.99e-15 ***
## lakenameTuesday Lake      -6.1413     0.6036 -10.174  < 2e-16 ***
## lakenameWard Lake         -2.5904     0.8139  -3.183  0.00146 ** 
## lakenameWest Long Lake    -5.3559     0.6124  -8.746  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 6.382 on 9494 degrees of freedom
## Multiple R-squared:  0.03014,    Adjusted R-squared:  0.02933 
## F-statistic: 36.88 on 8 and 9494 DF,  p-value: < 2.2e-16
```

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

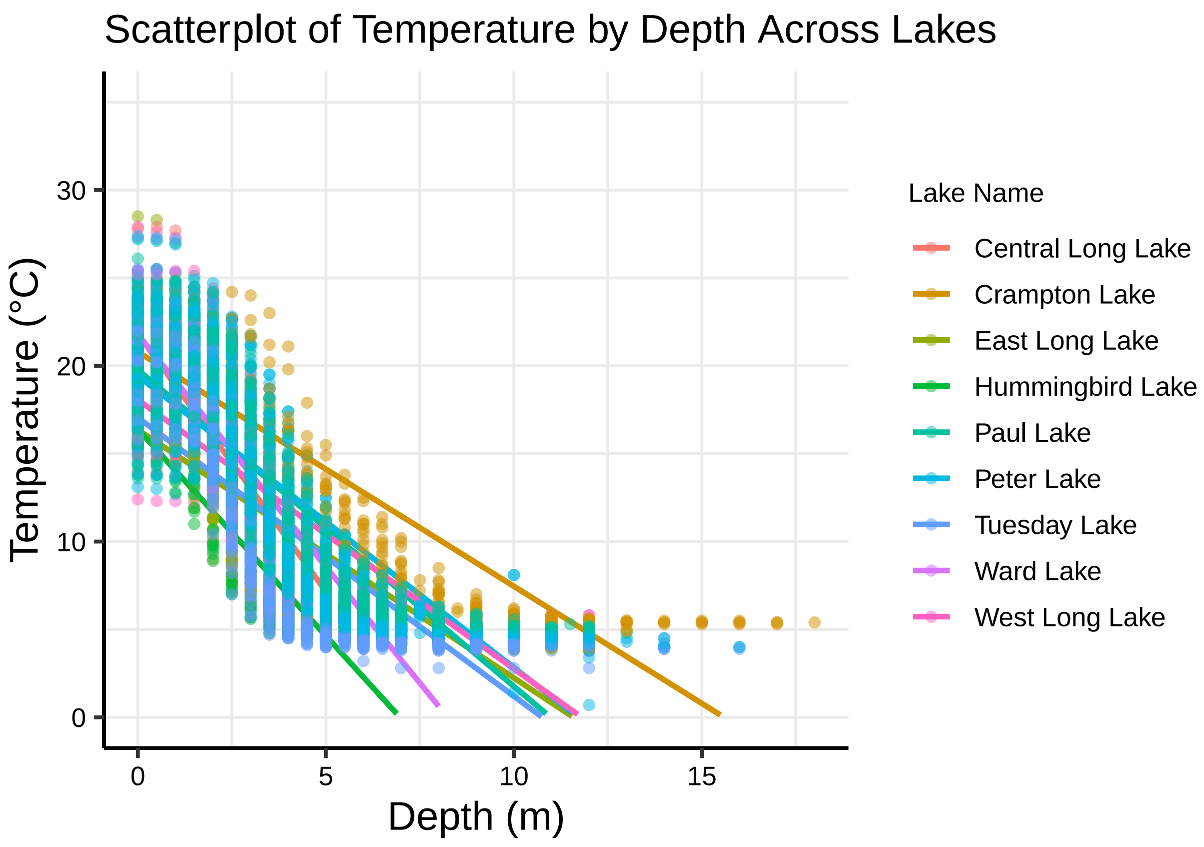
Answer: Yes, there is a significant difference in mean temperature among the lakes. From the ANOVA model, the p value is <2e-16, which is less than 0.05 and is statistically significant. Therefore, we can reject the null hypothesis that the mean of the lake temperature is the same across different lake sites, which means mean temperature is different among the lakes. From the linear model, we can see there is no coefficient associated with Central Long Lake. This is because its temperature mean is the base level and is represented by the intercept term. Meanwhile, the coefficient associated with all other lakes represent the difference of its mean temperature from mean temperature of Central Long Lake. From this linear model, each coefficient is statistically significant, since their p-value are less than 0.05. Therefore, from the linear model, we can also conclude that there is a significant difference in mean temperature among the lakes.

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 graph that depicts temperature by depth with a separate color for each lake
Temp.depth.scatt.color <- ggplot(NTL_LTER_Sub, aes(x = depth, y = temperature_C, color = lakename)) +
  geom_smooth (method="lm", se = FALSE)+
  geom_point(alpha = 0.5) +
  ylim(0, 35)+
  labs(x = "Depth (m)", y = "Temperature (°C)", color = "Lake Name") +
  labs(title = "Scatterplot of Temperature by Depth Across Lakes")
print (Temp.depth.scatt.color)
```

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

```
## Warning: Removed 88 rows containing missing values or values outside the scale range
## (`geom_smooth()`).
```



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

```
#15
#Using the Tukey's HSD test to determine which lakes have different means
TukeyHSD(Temp.anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = NTL_LTER_Sub)
##
## $lakename
##                                          diff         lwr         upr     p adj
## Crampton Lake-Central Long Lake    -2.8291387 -4.97533050 -0.68294692 0.0014395
## East Long Lake-Central Long Lake   -6.5010074 -8.41491151 -4.58710334 0.0000000
## Hummingbird Lake-Central Long Lake -6.7826598 -9.39740915 -4.16791044 0.0000000
## Paul Lake-Central Long Lake        -4.0856227 -5.92698164 -2.24426374 0.0000000
## Peter Lake-Central Long Lake       -4.5938464 -6.43239794 -2.75529489 0.0000000
## Tuesday Lake-Central Long Lake     -6.1412824 -8.01392896 -4.26863584 0.0000000
## Ward Lake-Central Long Lake        -2.5903736 -5.11555124 -0.06519587 0.0392367
## West Long Lake-Central Long Lake   -5.3558591 -7.25566616 -3.45605194 0.0000000
## East Long Lake-Crampton Lake       -3.6718687 -5.00938971 -2.33434772 0.0000000
## Hummingbird Lake-Crampton Lake     -3.9535211 -6.18126607 -1.72577609 0.0000013
## Paul Lake-Crampton Lake            -1.2564840 -2.48796128 -0.02500667 0.0413505
## Peter Lake-Crampton Lake           -1.7647077 -2.99198326 -0.53743215 0.0002831
## Tuesday Lake-Crampton Lake         -3.3121437 -4.58993033 -2.03435704 0.0000000
## Ward Lake-Crampton Lake             0.2387652 -1.88313397  2.36066428 0.9999938
## West Long Lake-Crampton Lake       -2.5267203 -3.84399049 -1.20945020 0.0000001
## Hummingbird Lake-East Long Lake    -0.2816524 -2.28658065  1.72327592 0.9999656
## Paul Lake-East Long Lake            2.4153847  1.65813576  3.17263372 0.0000000
## Peter Lake-East Long Lake           1.9071610  1.15676448  2.65755754 0.0000000
## Tuesday Lake-East Long Lake         0.3597250 -0.47071364  1.19016370 0.9180513
## Ward Lake-East Long Lake            3.9106339  2.02401108  5.79725666 0.0000000
## West Long Lake-East Long Lake       1.1451484  0.25515383  2.03514292 0.0021540
## Paul Lake-Hummingbird Lake          2.6970371  0.76123976  4.63283444 0.0005285
## Peter Lake-Hummingbird Lake         2.1888134  0.25568630  4.12194045 0.0132248
## Tuesday Lake-Hummingbird Lake       0.6413774 -1.32420489  2.60695968 0.9848376
## Ward Lake-Hummingbird Lake          4.1922862  1.59743925  6.78713323 0.0000193
## West Long Lake-Hummingbird Lake     1.4268007 -0.56467500  3.41827648 0.3907140
## Peter Lake-Paul Lake               -0.5082237 -1.04736090  0.03091344 0.0830294
## Tuesday Lake-Paul Lake             -2.0556597 -2.70157171 -1.40974770 0.0000000
## Ward Lake-Paul Lake                 1.4952491 -0.31773721  3.30823548 0.2044231
## West Long Lake-Paul Lake           -1.2702364 -1.99111377 -0.54935896 0.0000017
## Tuesday Lake-Peter Lake            -1.5474360 -2.18530058 -0.90957138 0.0000000
## Ward Lake-Peter Lake                2.0034729  0.19333794  3.81360778 0.0173703
## West Long Lake-Peter Lake          -0.7620126 -1.47568845 -0.04833682 0.0259997
## Ward Lake-Tuesday Lake              3.5509088  1.70615361  5.39566407 0.0000001
## West Long Lake-Tuesday Lake         0.7854233 -0.01198909  1.58283578 0.0573522
## West Long Lake-Ward Lake           -2.7654855 -4.63780592 -0.89316507 0.0001611
```

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: Paul Lake has the same mean temperature as Peter Lake has, statistically speaking. This is because the p-value of the mean temperature difference between Peter and Paul Lakes is 0.083, which is greater than 0.05. This result suggests that there is no statistical difference between the mean temperature of Peter Lake and Paul Lake. Yes, Central Long Lake has a mean temperature that is statistically distinct from all the other lakes. This is because for any lake combinition, Central Long Lake always has a statistically significant difference in mean temperature compared to other lake's mean temperatures.

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: The another test we might explore is a two-sample t-test. In this way, the null hypothesis will be Peter Lake and Paul Lake have the same mean temperatures. The alternative hypothesis will be Peter Lake and Paul Lake have different mean temperatures.

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?

```
#Wrangling the July data to include only records for Crampton Lake and Ward Lake
NTL_LTER_Sub_CW <- NTL_LTER_Sub%>%
  filter(lakename == "Crampton Lake" | lakename == "Ward Lake")
```

```r
#H0: There is no statistical significant difference between the mean temperature of
# Crampton Lake and Ward Lake. (Two samples have the same mean)
#Ha: There is statistical significant difference between the mean temperature of
# Crampton Lake and Ward Lake.(Two samples have different mean)

#Running the two-sample T-test on these data
C_W_twosample <- t.test(NTL_LTER_Sub_CW$temperature_C ~ NTL_LTER_Sub_CW$lakename)
C_W_twosample
```

```
##
##  Welch Two Sample t-test
##
## data:  NTL_LTER_Sub_CW$temperature_C by NTL_LTER_Sub_CW$lakename
## t = -0.35913, df = 242.64, p-value = 0.7198
## alternative hypothesis: true difference in means between group Crampton Lake and group Ward Lake is
## 95 percent confidence interval:
##  -1.548355  1.070825
## sample estimates:
## mean in group Crampton Lake     mean in group Ward Lake
##                    12.87003                    13.10880
```

> Answer: The t-test tells us that there is no statistically significant difference between the mean temperature of Crampton Lake and Ward Lake. This is because the p-value for this t-test is 0.7198, which is larger than 0.05. Therefore, we do not reject the null hypothesis, and this result suggests that Crampton Lake and Ward Lake have equal mean July temperature. This result also matches with the answer for part 16 because the p-value of the Tukey's HSD test for the Ward Lake-Crampton Lake pair is 0.999 and greater than 0.05, which also suggests that there is no statistical difference between the mean temperature of Crampton Lake and Ward Lake.