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

Danielle Butler

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.

```r
#1
#Import libraries
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   3.5.1     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
```

```r
library(lubridate)
library(here)
```

```
## here() starts at /home/guest/EDA_Spring2025
```

```r
library(knitr)
library(agricolae)
library(ggplot2)

here()
```

```
## [1] "/home/guest/EDA_Spring2025"
```

```r
#import data
NTL_LTER_Raw <- read.csv(here("Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv"), stringsAsFactors = TRU

#Set date to date format
NTL_LTER_Raw$sampledate <- as.Date(NTL_LTER_Raw$sampledate,format = "%m/%d/%y")
class(NTL_LTER_Raw$sampledate)
```

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

```r
#2
#Set theme
mytheme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
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: The mean lake temperature recorded during July does not change with depth across all lakes. Ha: The mean lake temperature recorded during July changes with depth across all lakes.

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.

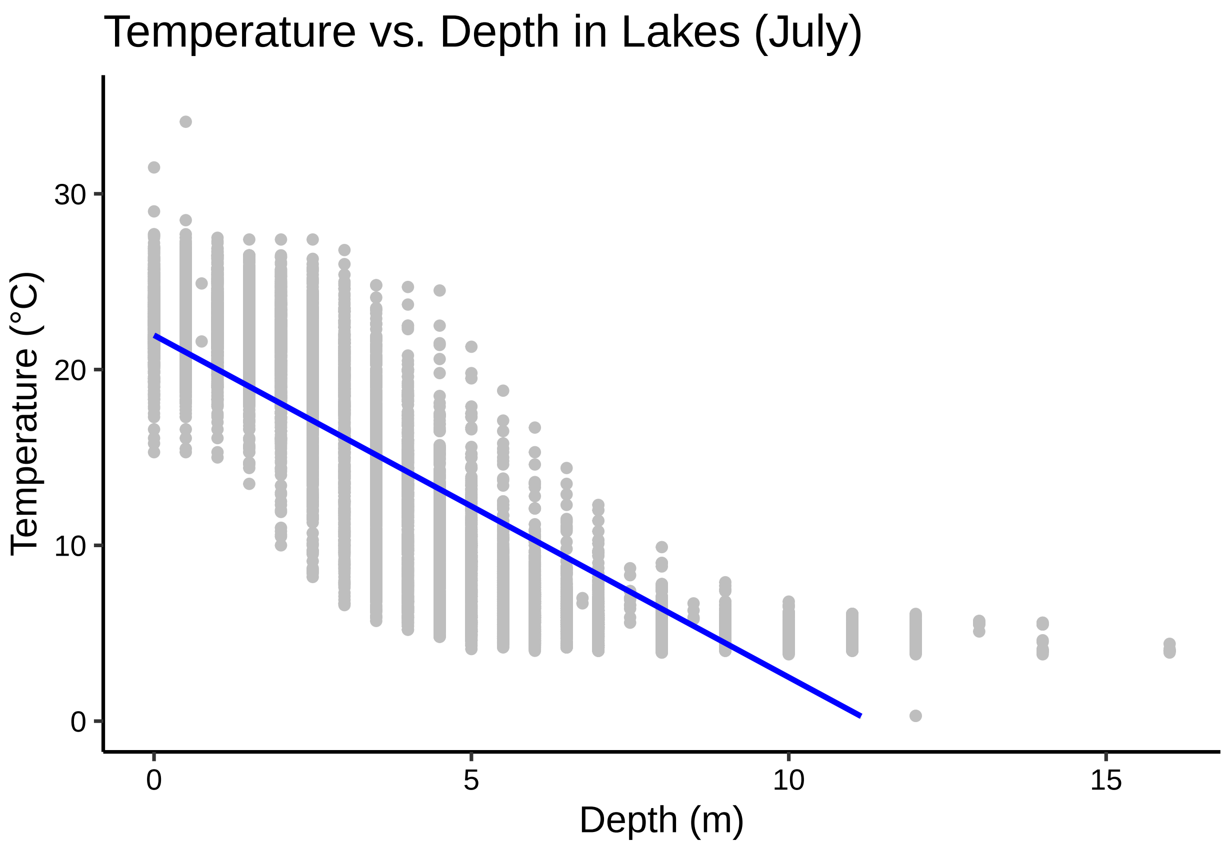
```r
#4
#Wrangle the data
NTL_LTER_Wrangled <- NTL_LTER_Raw %>%
  filter(format(sampledate,"%m") == "07") %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  na.omit()
summary(NTL_LTER_Wrangled)
```

```
##            lakename        year4          daynum          depth
##  Peter Lake    :2872   Min.   :1984   Min.   :182.0   Min.   : 0.000
##  Paul Lake     :2660   1st Qu.:1992   1st Qu.:190.0   1st Qu.: 2.000
##  Tuesday Lake  :1524   Median :1998   Median :198.0   Median : 4.500
##  West Long Lake:1026   Mean   :1999   Mean   :197.5   Mean   : 4.745
##  East Long Lake: 968   3rd Qu.:2006   3rd Qu.:205.0   3rd Qu.: 7.000
##  Crampton Lake : 318   Max.   :2016   Max.   :213.0   Max.   :16.000
##  (Other)       : 360
##  temperature_C
##  Min.   : 0.30
##  1st Qu.: 5.50
##  Median :10.10
##  Mean   :12.72
##  3rd Qu.:20.80
##  Max.   :34.10
##
```

```r
#5
ggplot(NTL_LTER_Wrangled, aes(x = depth, y = temperature_C)) +
  geom_point(color = "gray") +
  geom_smooth(method = "lm", color = "blue") +
  scale_y_continuous(limits = c(0, 35)) +
  labs(title = "Temperature vs. Depth in Lakes (July)",
       x = "Depth (m)",
       y = "Temperature (°C)")
```

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

```
## Warning: Removed 24 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: The deeper the depth in meters, the lower the temperature in C. While the smooth line is linear in nature, the distribution of points may point out a not perfect linear relationship.

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

```
#7
NTL_LTER.regression <-
  lm(NTL_LTER_Wrangled$depth ~
       NTL_LTER_Wrangled$temperature_C)

summary(NTL_LTER.regression)
```

```
##
## Call:
## lm(formula = NTL_LTER_Wrangled$depth ~ NTL_LTER_Wrangled$temperature_C)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0685 -1.1065 -0.2334  0.9668  8.0964
##
## Coefficients:
```

```
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      9.573728   0.033803   283.2   <2e-16 ***
## NTL_LTER_Wrangled$temperature_C -0.379578   0.002289  -165.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.694 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
```

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 coefficient estimate of -0.379578 suggests the negative correlat- ion between depth and temperature. As depth increases, temperature decreases. The multiple R squared of 0.7387 suggests a strong relationship between the two variables.The p value of <2.2e-16 also suggests a strong relationship between the two variables and is highly significant. There are 9726 degrees of freedom, which is a large sample size. For every 1m increase in depth, temperature is predicted to decrease by approximately 2.63°C.

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
NTL_LTER_Wrangled_AIC <- lm(data = NTL_LTER_Wrangled, temperature_C ~ year4
                            + daynum + depth)

step(NTL_LTER_Wrangled_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 = NTL_LTER_Wrangled)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##    -8.57556      0.01134      0.03978     -1.94644
```

```r
model <- lm(data = NTL_LTER_Wrangled, temperature_C ~ year4
                            + daynum + depth)
summary(model)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL_LTER_Wrangled)
##
## 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
```

```r
#10
NTL_LTER_Wrangled_MR <- lm(data = subset(NTL_LTER_Wrangled),
                   temperature_C ~ year4 + daynum + depth)

summary(NTL_LTER_Wrangled_MR)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = subset(NTL_LTER_Wrangled))
##
## 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: It suggests using all 3 variables to predict temperature. As they all have significant p values. The observed variance shows that 74% of the variance in temperature is explained. The multiple R squared is very similar in the first linear model at 0.7387 when we only used depth as the explanatory variable. So there is barely an improvement between using the 1 variable of depth versus add- ing these other two to predict temperature.

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

NTL_LTER.anova <- aov(data = NTL_LTER_Wrangled, depth ~ temperature_C)
summary(NTL_LTER.anova)
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## temperature_C    1  78877   78877   27501 <2e-16 ***
## Residuals     9726  27896       3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
NTL_LTER.anova2 <- lm(data = NTL_LTER_Wrangled, depth ~ temperature_C)
summary(NTL_LTER.anova2)
```

```
##
## Call:
## lm(formula = depth ~ temperature_C, data = NTL_LTER_Wrangled)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0685 -1.1065 -0.2334  0.9668  8.0964
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.573728   0.033803   283.2   <2e-16 ***
```

```
## temperature_C -0.379578   0.002289  -165.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.694 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
```

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 due to the small p value provided in the ANOVA.

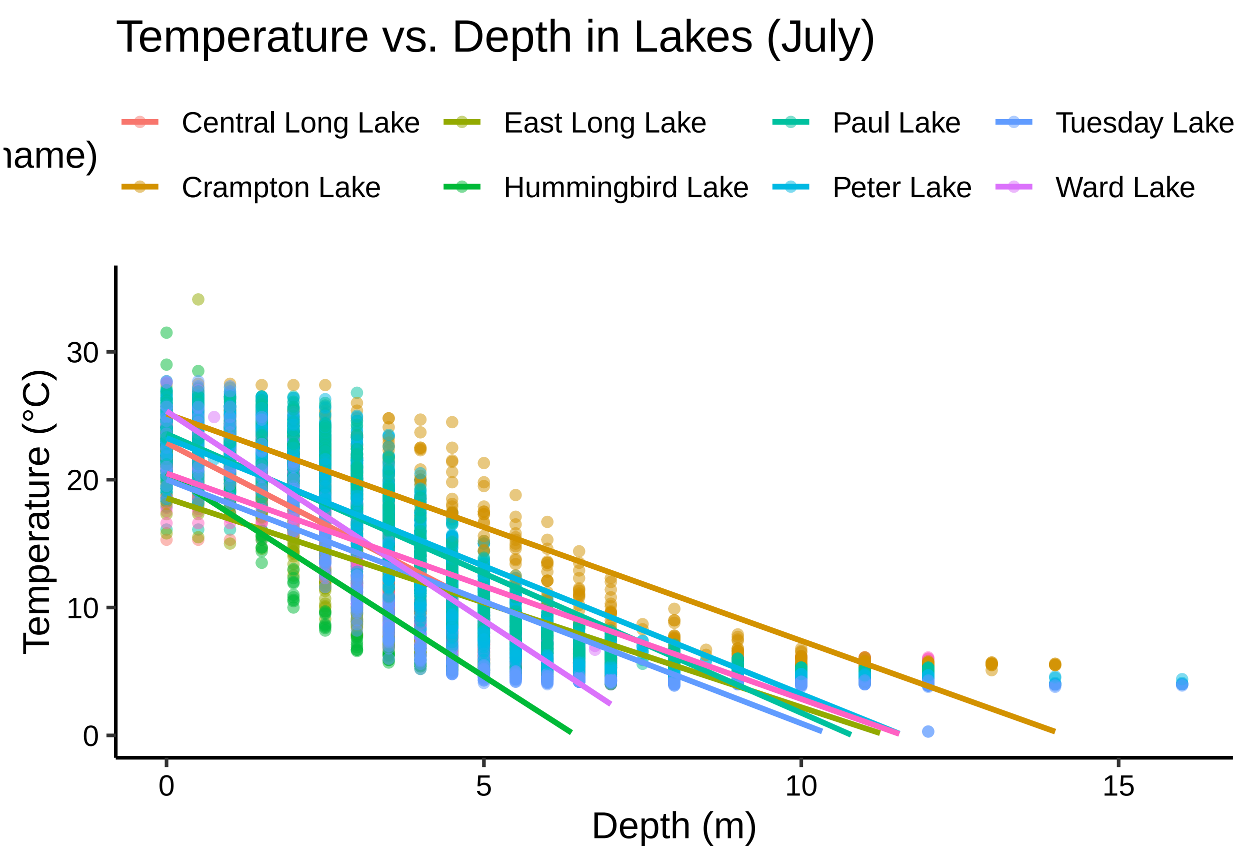
14. Create a graph that depicts temperature by depth, with a separate color for each lake. Add a geom_smooth (method = "lm", se = FALSE) for each lake. Make your points 50 % transparent. Adjust your y axis limits to go from 0 to 35 degrees. Clean up your graph to make it pretty.

```r
#14.
#Graph the results
plot <- ggplot(NTL_LTER_Wrangled, aes(x = depth, y = temperature_C, color = factor(lakename))) +
  geom_point(alpha=0.5) +
  geom_smooth(method = "lm", se=FALSE,aes(group = factor(lakename))) +
  scale_y_continuous(limits = c(0, 35)) +
  labs(title = "Temperature vs. Depth in Lakes (July)",
       x = "Depth (m)",
       y = "Temperature (°C)")

print(plot)
```

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

```
## Warning: Removed 73 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
# Format as aov
NTL_LTER_2way <- aov(data = NTL_LTER_Wrangled, temperature_C ~factor(lakename))
summary(NTL_LTER_2way)
```

```
##                     Df Sum Sq Mean Sq F value Pr(>F)
## factor(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
```

```
# Format as lm
NTL_LTER_2way2 <- lm(data = NTL_LTER_Wrangled, temperature_C ~factor(lakename))
summary(NTL_LTER_2way2)
```

```
##
## Call:
## lm(formula = temperature_C ~ factor(lakename), data = NTL_LTER_Wrangled)
##
## 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 ***
## factor(lakename)Crampton Lake    -2.3145     0.7699  -3.006 0.002653 **
## factor(lakename)East Long Lake   -7.3987     0.6918 -10.695  < 2e-16 ***
## factor(lakename)Hummingbird Lake -6.8931     0.9429  -7.311 2.87e-13 ***
## factor(lakename)Paul Lake        -3.8522     0.6656  -5.788 7.36e-09 ***
## factor(lakename)Peter Lake       -4.3501     0.6645  -6.547 6.17e-11 ***
## factor(lakename)Tuesday Lake     -6.5972     0.6769  -9.746  < 2e-16 ***
## factor(lakename)Ward Lake        -3.2078     0.9429  -3.402 0.000672 ***
## factor(lakename)West 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
```

```
# Run a post-hoc test for pairwise differences
TukeyHSD(NTL_LTER_2way)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ factor(lakename), data = NTL_LTER_Wrangled)
##
## $`factor(lakename)`
##                                        diff        lwr        upr     p adj
## Crampton Lake-Central Long Lake    -2.3145195 -4.7031913  0.0741524 0.0661566
## East Long Lake-Central Long Lake   -7.3987410 -9.5449411 -5.2525408 0.0000000
## Hummingbird Lake-Central Long Lake -6.8931304 -9.8184178 -3.9678430 0.0000000
## Paul Lake-Central Long Lake        -3.8521506 -5.9170942 -1.7872070 0.0000003
## Peter Lake-Central Long Lake       -4.3501458 -6.4115874 -2.2887042 0.0000000
## Tuesday Lake-Central Long Lake     -6.5971805 -8.6971605 -4.4972005 0.0000000
## Ward Lake-Central Long Lake        -3.2077856 -6.1330730 -0.2824982 0.0193405
## West Long Lake-Central Long Lake   -6.0877513 -8.2268550 -3.9486475 0.0000000
## East Long Lake-Crampton Lake       -5.0842215 -6.5591700 -3.6092730 0.0000000
## Hummingbird Lake-Crampton Lake     -4.5786109 -7.0538088 -2.1034131 0.0000004
## Paul Lake-Crampton Lake            -1.5376312 -2.8916215 -0.1836408 0.0127491
## Peter Lake-Crampton Lake           -2.0356263 -3.3842699 -0.6869828 0.0000999
## Tuesday Lake-Crampton Lake         -4.2826611 -5.6895065 -2.8758157 0.0000000
## Ward Lake-Crampton Lake            -0.8932661 -3.3684639  1.5819317 0.9714459
## West Long Lake-Crampton Lake       -3.7732318 -5.2378351 -2.3086285 0.0000000
## Hummingbird Lake-East Long Lake     0.5056106 -1.7364925  2.7477137 0.9988050
## Paul Lake-East Long Lake            3.5465903  2.6900206  4.4031601 0.0000000
## Peter Lake-East Long Lake           3.0485952  2.2005025  3.8966879 0.0000000
## Tuesday Lake-East Long Lake         0.8015604 -0.1363286  1.7394495 0.1657485
## Ward Lake-East Long Lake            4.1909554  1.9488523  6.4330585 0.0000002
## West Long Lake-East Long Lake       1.3109897  0.2885003  2.3334791 0.0022805
## Paul Lake-Hummingbird Lake          3.0409798  0.8765299  5.2054296 0.0004495
## Peter Lake-Hummingbird Lake         2.5429846  0.3818755  4.7040937 0.0080666
## Tuesday Lake-Hummingbird Lake       0.2959499 -1.9019508  2.4938505 0.9999752
## Ward Lake-Hummingbird Lake          3.6853448  0.6889874  6.6817022 0.0043297
```

```
## West Long Lake-Hummingbird Lake     0.8053791 -1.4299320  3.0406903 0.9717297
## Peter Lake-Paul Lake               -0.4979952 -1.1120620  0.1160717 0.2241586
## Tuesday Lake-Paul Lake             -2.7450299 -3.4781416 -2.0119182 0.0000000
## Ward Lake-Paul Lake                 0.6443651 -1.5200848  2.8088149 0.9916978
## West Long Lake-Paul Lake           -2.2356007 -3.0742314 -1.3969699 0.0000000
## Tuesday Lake-Peter Lake            -2.2470347 -2.9702236 -1.5238458 0.0000000
## Ward Lake-Peter Lake                1.1423602 -1.0187489  3.3034693 0.7827037
## West Long Lake-Peter Lake          -1.7376055 -2.5675759 -0.9076350 0.0000000
## Ward Lake-Tuesday Lake              3.3893950  1.1914943  5.5872956 0.0000609
## West Long Lake-Tuesday Lake         0.5094292 -0.4121051  1.4309636 0.7374387
## West Long Lake-Ward Lake           -2.8799657 -5.1152769 -0.6446546 0.0021080
```

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 & Ward Lake have the same mean temperature as Peter Lake based on the p values. Central Long Lake looks to have a distinct mean temp from all 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: You could do a two sample t test if you were only looking at those 2 lakes.

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?

```
#Wrangle to two lakes
NTL_LTER_Two <- NTL_LTER_Wrangled %>%
 filter(lakename %in% c('Crampton Lake', 'Ward Lake')) %>%
  filter(complete.cases(.))
summary(NTL_LTER_Two)
```

```
##               lakename       year4          daynum          depth
##  Crampton Lake    :318   Min.   :1999   Min.   :183.0   Min.   : 0.000
##  Ward Lake        :116   1st Qu.:2004   1st Qu.:188.0   1st Qu.: 2.000
##  Central Long Lake:  0   Median :2005   Median :197.0   Median : 4.500
##  East Long Lake   :  0   Mean   :2006   Mean   :196.7   Mean   : 4.937
##  Hummingbird Lake :  0   3rd Qu.:2010   3rd Qu.:204.0   3rd Qu.: 7.000
##  Paul Lake        :  0   Max.   :2012   Max.   :211.0   Max.   :14.000
##  (Other)          :  0
##  temperature_C
##  Min.   : 5.00
##  1st Qu.: 7.40
##  Median :15.30
##  Mean   :15.11
##  3rd Qu.:22.38
##  Max.   :27.60
##
```

```r
unique(NTL_LTER_Two$lakename)
```

```
## [1] Crampton Lake Ward Lake
## 9 Levels: Central Long Lake Crampton Lake East Long Lake ... West Long Lake
```

```r
#Format as a t-test
t_test_result <- t.test(temperature_C ~ lakename, data = NTL_LTER_Two)
t_test_result
```

```
##
##  Welch Two Sample t-test
##
## data:  temperature_C by lakename
## t = 1.1181, df = 200.37, p-value = 0.2649
## alternative hypothesis: true difference in means between group Crampton Lake and group Ward Lake is n
## 95 percent confidence interval:
##  -0.6821129  2.4686451
## sample estimates:
## mean in group Crampton Lake     mean in group Ward Lake
##                    15.35189                    14.45862
```

> Answer: There is not enough data to say that the two lakes have a different temperature in July. The two lakes average temperatures are very similar between 15.35 and 14.45 C. This does not match my answer above.