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

Quinn Bankson

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
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.2.1     v dplyr   1.1.2
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.4     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

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

```
##
## Attaching package: 'lubridate'
##
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(dplyr)
library(cowplot)
```

```
##
## Attaching package: 'cowplot'
##
## The following object is masked from 'package:lubridate':
##
##     stamp
```

```
library(ggplot2)
library(ggrepel)
library(here)
```

```
## here() starts at /Users/mac/Documents/EDE_Fall2023
```

```
ntllter <-read.csv((file = "/Users/mac/Documents/EDE_Fall2023/Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Ra

#2

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: There is no relationship between mean lake temperature in July and depth across all lakes in July. Ha: The depth of the lake affects the mean lake temperature in July across all lakes.

4. Wrangle your NTL-LTER dataset with a pipe function so that the records meet the following criteria:

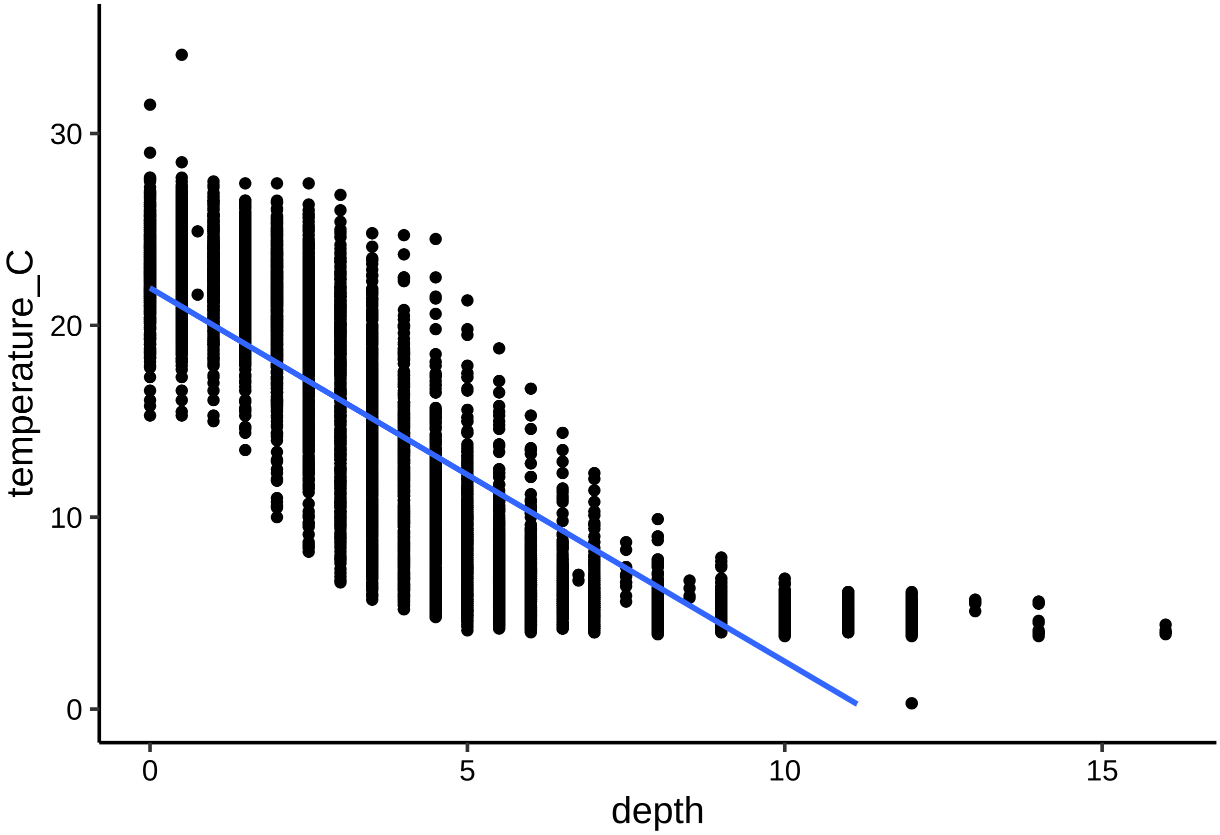
- Only dates in July.
- Only the columns: `lakename`, `year4`, `daynum`, `depth`, `temperature_C`
- Only complete cases (i.e., remove NAs)

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.

```
#4

july_df <- ntllter %>%
  filter(daynum >= 182, daynum <= 212) %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  na.omit()




#5
july_scatter <- ggplot(july_df, aes(x=depth, y = temperature_C)) +
    geom_point() +
    ylim(0, 35) +
    geom_smooth(method = "lm")

print(july_scatter)
```

```
## `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 figure suggests that a lower depth is associated with a lower temperature.

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

```
#7
ltr_reg <- lm(
  data = july_df,
  temperature_C ~ depth)
summary(ltr_reg)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = july_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5077 -3.0182  0.0743  2.9248 13.6033
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.94872    0.06790   323.3   <2e-16 ***
## depth       -1.94700    0.01173  -166.0   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.829 on 9720 degrees of freedom
## Multiple R-squared:  0.7391, Adjusted R-squared:  0.7391
## F-statistic: 2.754e+04 on 1 and 9720 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: COEFFICIENT: Temperature is predicted to decrease by an average of 1.94 degrees C for a change of 1m in depth. These results are statistically significant beyond the 99.9% confidence level. INTERCEPT: Temperature is predicted to have a temperature of 21.95 C at a depth of 0 m. These results are statistically significant beyond the 99.9% confidence level. About 73.9 percent of the variance in temperature is explained only by changes in depth. DF = 9720

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.

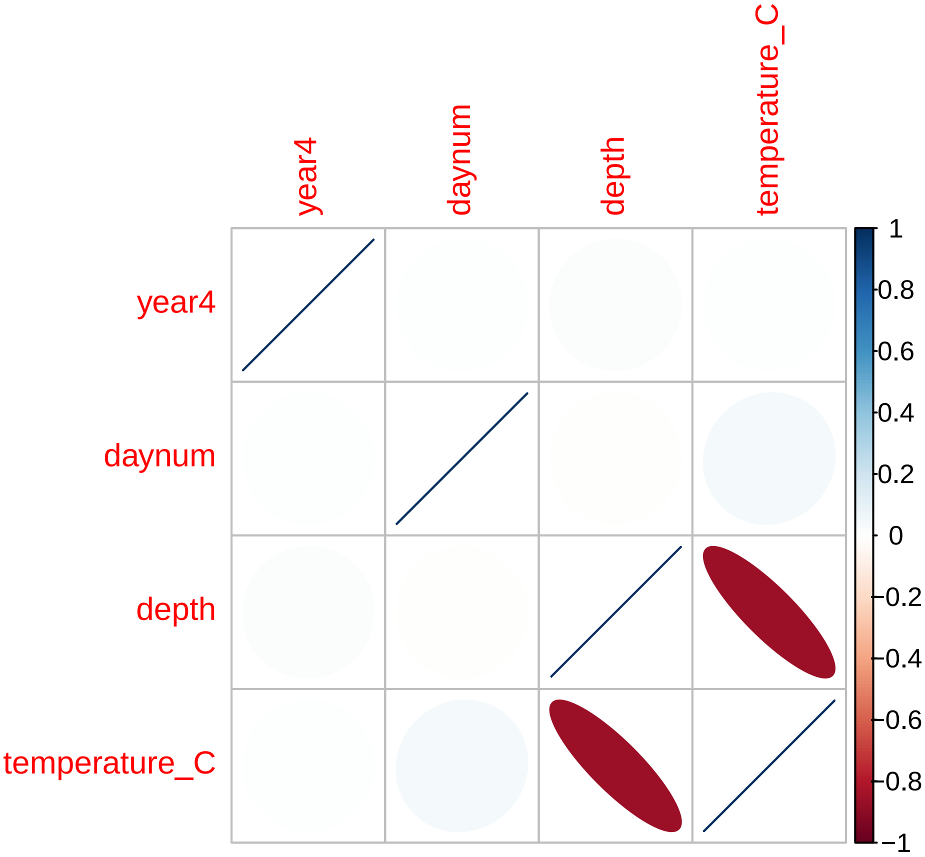
9. Run an AIC to determine what set of explanatory variables (year4, daynum, depth) is best suited to predict temperature.

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

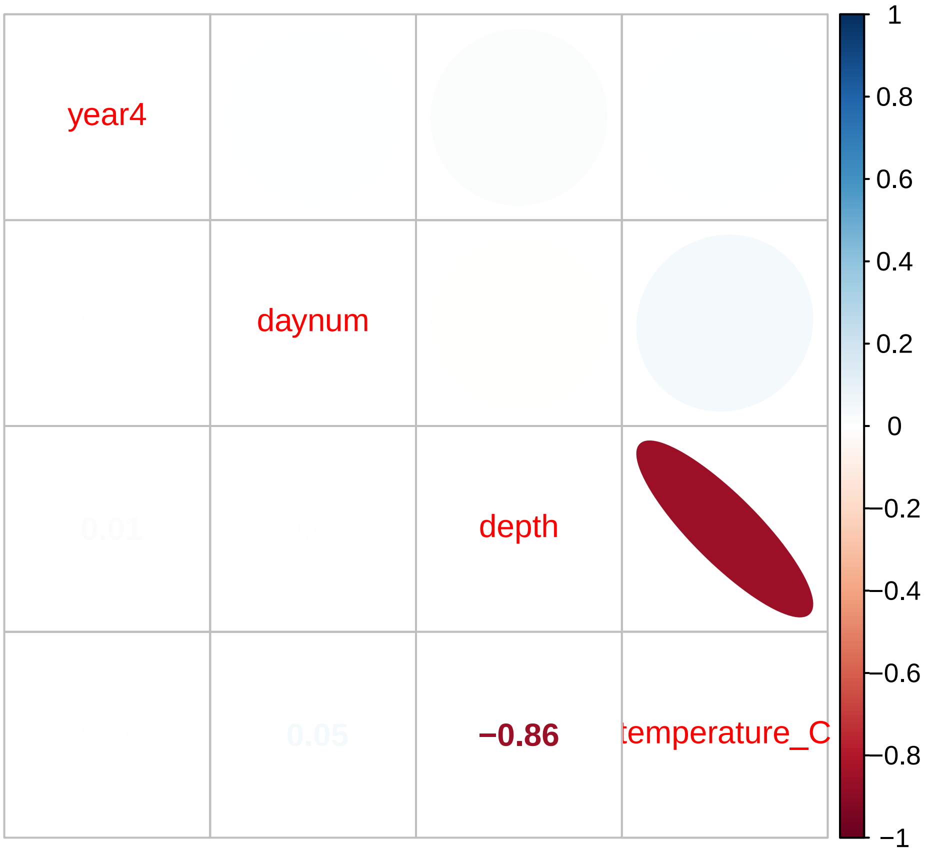
```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
july_df_sub <- july_df %>%
  mutate(across(everything(), as.numeric)) %>%
  select(year4, daynum, depth, temperature_C)
july_corr_sub <- cor(july_df_sub)
corrplot(july_corr_sub, method = "ellipse")
```



```
corrplot.mixed(july_corr_sub, upper = "ellipse")
```



```r
#9
tempbyALL.regression <- lm(data = july_df,
                           temperature_C ~ year4 + daynum + depth)
step(tempbyALL.regression)
```

```
## Start:  AIC=26016.31
## temperature_C ~ year4 + daynum + depth
##
##            Df Sum of Sq    RSS   AIC
## <none>                  141118 26016
## - year4    1        80 141198 26020
## - daynum   1      1333 142450 26106
## - depth    1    403925 545042 39151
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = july_df)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##    -6.45556      0.01013      0.04134     -1.94726
```

```r
#10
#AIC indicates that removing NONE of the predictors will improve the model.
```

```
temp_multivar.regression <- lm(data = july_df,
                           temperature_C ~ year4 + daynum + depth)
summary(temp_multivar.regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = july_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6517 -2.9937  0.0855  2.9692 13.6171
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -6.455560   8.638808   -0.747   0.4549
## year4        0.010131   0.004303    2.354   0.0186 *
## daynum       0.041336   0.004315    9.580   <2e-16 ***
## depth       -1.947264   0.011676 -166.782   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.811 on 9718 degrees of freedom
## Multiple R-squared:  0.7417, Adjusted R-squared:  0.7417
## F-statistic:  9303 on 3 and 9718 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: #AIC indicates that removing NONE of the predictors will improve the model. The model is strongest with year4, daynum, and depth. The multivariable model has an rsquared of 0.7417, which is higher than the single variable 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
july.anova <- aov(data = july_df, temperature_C ~ lakename)
summary(july.anova)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## lakename       8  21214  2651.8   49.04 <2e-16 ***
## Residuals   9713 525188    54.1
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
july.anova2 <- lm(data = july_df, temperature_C ~ lakename)
summary(july.anova2)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = july_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.766  -6.592  -2.692   7.634  23.832
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               17.6731     0.6741  26.218  < 2e-16 ***
## lakenameCrampton Lake     -2.3212     0.7902  -2.938  0.00332 **
## lakenameEast Long Lake    -7.4054     0.7143 -10.367  < 2e-16 ***
## lakenameHummingbird Lake  -6.8998     0.9594  -7.192 6.88e-13 ***
## lakenamePaul Lake         -3.8813     0.6891  -5.633 1.82e-08 ***
## lakenamePeter Lake        -4.3710     0.6878  -6.355 2.18e-10 ***
## lakenameTuesday Lake      -6.6073     0.7002  -9.437  < 2e-16 ***
## lakenameWard Lake         -3.2145     0.9594  -3.350  0.00081 ***
## lakenameWest Long Lake    -6.0876     0.7115  -8.556  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.353 on 9713 degrees of freedom
## Multiple R-squared:  0.03883,    Adjusted R-squared:  0.03803
## F-statistic: 49.04 on 8 and 9713 DF,  p-value: < 2.2e-16
```

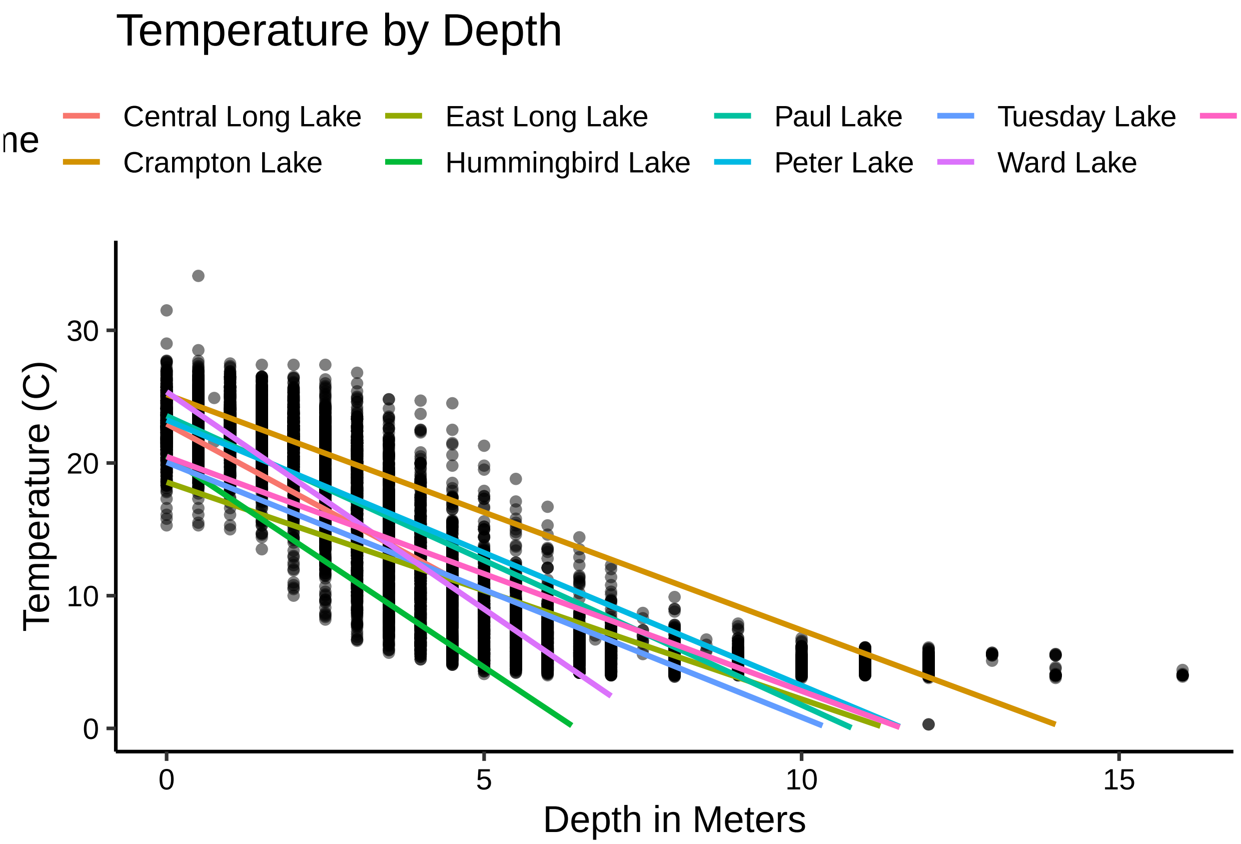
13. Is there a significant difference in mean temperature among the lakes? Report your findings.

    Answer: There are significant differences in mean temperature among the lakes. The ANOVA testing reveals the temperatures grouped by lakename have significantly different means. Three astriks are displayed, letting us know that the differences in mean are significant above the .99 level. The LM model shows similar results. Nearly every lakename group has three astriks by its coefficient, indicating a high amount of significance in the different means.

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.

tempby_name_depth <- ggplot(july_df, aes(x=depth, y=temperature_C)) +
  geom_point(alpha = 0.5) +
  geom_smooth(aes(color=lakename), method = "lm", se = FALSE) +
  ylim(0,35) +
  labs(title = "Temperature by Depth", x = "Depth in Meters", y = "Temperature (C)")

print(tempby_name_depth)
```

```
## `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
TukeyHSD(july.anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = july_df)
##
## $lakename
##                                          diff        lwr        upr     p adj
## Crampton Lake-Central Long Lake    -2.3212225 -4.7727515  0.1303066 0.0801309
## East Long Lake-Central Long Lake   -7.4054440 -9.6215318 -5.1893561 0.0000000
## Hummingbird Lake-Central Long Lake -6.8998334 -9.8763946 -3.9232722 0.0000000
## Paul Lake-Central Long Lake        -3.8813120 -6.0191419 -1.7434822 0.0000007
## Peter Lake-Central Long Lake       -4.3710346 -6.5048955 -2.2371736 0.0000000
## Tuesday Lake-Central Long Lake     -6.6072831 -8.7795517 -4.4350145 0.0000000
## Ward Lake-Central Long Lake        -3.2144886 -6.1910498 -0.2379273 0.0229685
## West Long Lake-Central Long Lake   -6.0875867 -8.2949346 -3.8802388 0.0000000
## East Long Lake-Crampton Lake       -5.0842215 -6.5587481 -3.6096949 0.0000000
## Hummingbird Lake-Crampton Lake     -4.5786109 -7.0531008 -2.1041211 0.0000003
## Paul Lake-Crampton Lake            -1.5600896 -2.9141574 -0.2060217 0.0106305
## Peter Lake-Crampton Lake           -2.0498121 -3.3976050 -0.7020192 0.0000841
## Tuesday Lake-Crampton Lake         -4.2860606 -5.6938725 -2.8782488 0.0000000
## Ward Lake-Crampton Lake            -0.8932661 -3.3677559  1.5812237 0.9713958
## West Long Lake-Crampton Lake       -3.7663643 -5.2277226 -2.3050060 0.0000000
## Hummingbird Lake-East Long Lake     0.5056106 -1.7358512  2.7470723 0.9988025
## Paul Lake-East Long Lake            3.5241319  2.6670727  4.3811912 0.0000000
## Peter Lake-East Long Lake           3.0344094  2.1872987  3.8815201 0.0000000
## Tuesday Lake-East Long Lake         0.7981609 -0.1415120  1.7378337 0.1721160
## Ward Lake-East Long Lake            4.1909554  1.9494937  6.4324171 0.0000002
## West Long Lake-East Long Lake       1.3178572  0.2997124  2.3360021 0.0019544
## Paul Lake-Hummingbird Lake          3.0185213  0.8543999  5.1826428 0.0005172
## Peter Lake-Hummingbird Lake         2.5287988  0.3685979  4.6889997 0.0086420
## Tuesday Lake-Hummingbird Lake       0.2925503 -1.9055981  2.4906986 0.9999773
## Ward Lake-Hummingbird Lake          3.6853448  0.6898445  6.6808451 0.0043115
## West Long Lake-Hummingbird Lake     0.8122467 -1.4205745  3.0450678 0.9700210
## Peter Lake-Paul Lake               -0.4897225 -1.1036180  0.1241730 0.2442990
## Tuesday Lake-Paul Lake             -2.7259711 -3.4623514 -1.9895907 0.0000000
## Ward Lake-Paul Lake                 0.6668235 -1.4972980  2.8309450 0.9895659
## West Long Lake-Paul Lake           -2.2062747 -3.0404749 -1.3720745 0.0000000
## Tuesday Lake-Peter Lake            -2.2362485 -2.9610258 -1.5114713 0.0000000
## Ward Lake-Peter Lake                1.1565460 -1.0036549  3.3167469 0.7703831
## West Long Lake-Peter Lake          -1.7165522 -2.5405279 -0.8925764 0.0000000
## Ward Lake-Tuesday Lake              3.3927945  1.1946462  5.5909429 0.0000597
## West Long Lake-Tuesday Lake         0.5196964 -0.3991749  1.4385677 0.7121762
## West Long Lake-Ward Lake           -2.8730982 -5.1059193 -0.6402770 0.0021521
```

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: These lakes have the same mean temperature as Peter Lake, statistically speaking: Crampton Lake, Hummingbird Lake, West Long Lake, Ward Lake. The lakes differences have a p value that suggests we cannot reject the null hypothesis (Null: there is not a difference). Central Long lake, East Long lake, Paul lake, and Tuesday lake do not have pvalues of statistical significance with any other lakes. This means they are statistically distinct from 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: We might use a Hausman test. Hausman tests are useful when deciding on research methods with panel data such as pooled OLS, Fixed Effects, or Random Effects. They can also be used in a more basic sense (like a diff in diff or a regression) to determine if coefficients are different. I am more familiar with the Hausman command on STATA but I am sure there is an R 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?

```
julyCram.Ward <- july_df %>%
  filter(lakename =="Crampton Lake" | lakename == "Ward Lake")

tw.ttest <- lm(julyCram.Ward$temperature_C ~ julyCram.Ward$lakename)
summary(tw.ttest)
```

```
##
## Call:
## lm(formula = julyCram.Ward$temperature_C ~ julyCram.Ward$lakename)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3519  -7.5286   0.1947   7.0481  13.1414
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       15.3519     0.4087   37.56   <2e-16 ***
## julyCram.Ward$lakenameWard Lake  -0.8933     0.7906   -1.13    0.259
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.289 on 432 degrees of freedom
## Multiple R-squared:  0.002946,   Adjusted R-squared:  0.0006383
## F-statistic: 1.277 on 1 and 432 DF,  p-value: 0.2592
```

> Answer: The coefficient for temperature is not statistically significant. We fail to reject the null hypothesis, which means the values are not statistically significant. The TukeyHSD test indicated that the values are not statistically significant different either.