Assignment 6: GLMs (Linear Regressios, ANOVA, & t-tests)

Tasha Griffiths

OVERVIEW

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

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Fay_A06_GLMs.Rmd") prior to submission.

The completed exercise is due on Monday, February 28 at 7:00 pm.

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 set up session
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

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

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

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

```
getwd()
```

```
## [1] "C:/Users/Tasha Griffiths/Documents/Duke Year 1/Spring 22 Classes/Environmental Data Analytics/G
```

```
NTL.LTER <- read.csv("./Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv",
                     stringsAsFactors = TRUE)

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

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

```
#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?

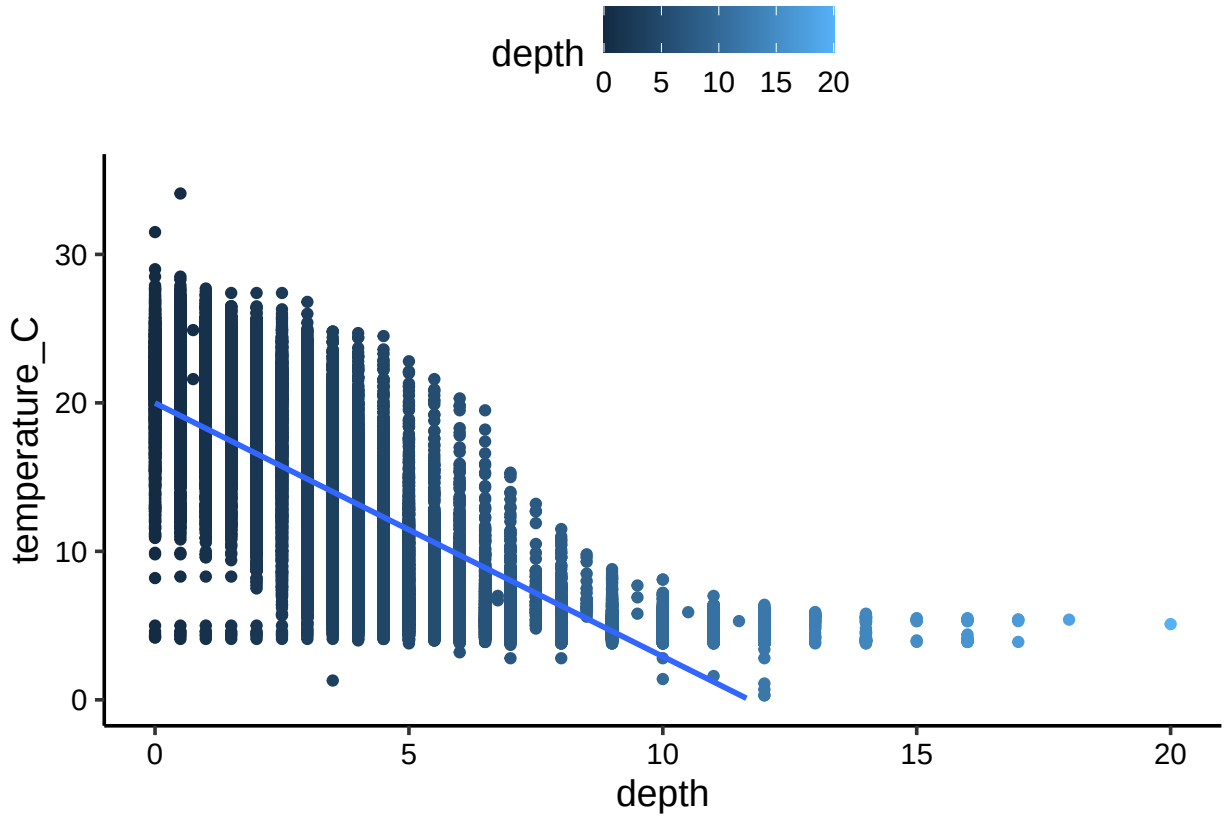
3. State the null and alternative hypotheses for this question: > Answer: H0: Null Hypothesis - Lake temperature in July does not change with depth across all lakes Ha: Alternative Hypothesis - Lake temperature in July does change with depth across all lakes
4. Wrangle your NTL-LTER dataset with a pipe 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.

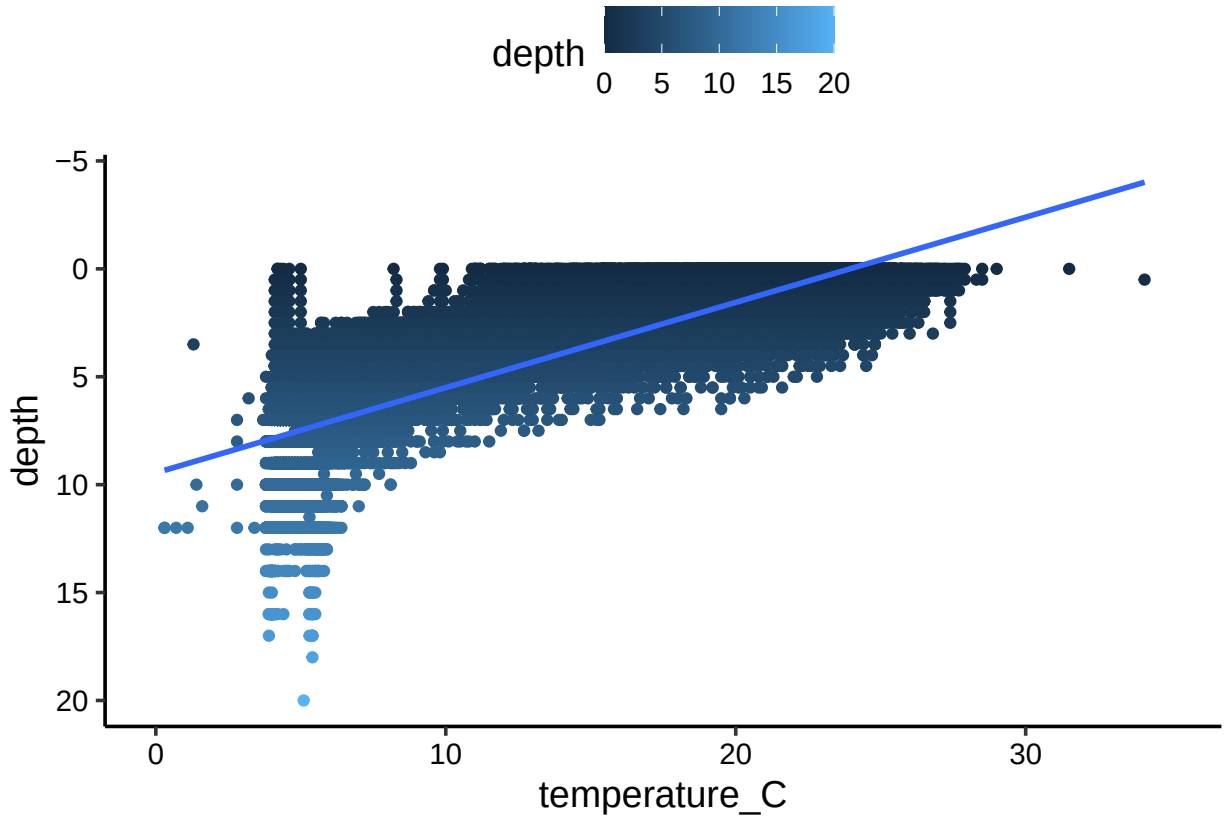
```r
#4 wrangle data
NTL.LTER.JUL <- NTL.LTER %>% mutate(NTL.LTER, month = month(sampledate)) %>%
  select("lakename", "year4", "daynum", "depth", "temperature_C") %>%
  na.omit()


#5 visualize with scatter plot
scatter.temp.depth <- ggplot(NTL.LTER.JUL, aes(x = depth , y = temperature_C,
                                               col = depth)) +
  geom_point() + geom_smooth(method = lm) + ylim(0, 35)

print(scatter.temp.depth)
```

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

## Warning: Removed 33 rows containing missing values (geom_smooth).
```




```r
#flipped axis to visualize easier
flipped.scatter.temp.depth <- ggplot(NTL.LTER.JUL, aes(x = temperature_C ,
                                                       y = depth,
                                                       col = depth)) +
  geom_point() + geom_smooth(method = lm) + xlim(0, 35) + scale_y_reverse()

print(flipped.scatter.temp.depth)
```

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



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 scatter plot suggests that temperature does decrease with depth. If we were just looking at this plot it would appear we can reject the null hypothesis. The slope of the lm line does appear to be different than zero sugguesting a linear relationship.

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

```
#7 linear regression on temp and depth
temp.regression <- lm(data = NTL.LTER.JUL, temperature_C ~ depth)
summary(temp.regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = NTL.LTER.JUL)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7864  -3.1363  -0.1219   3.1815  19.2568
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.986395   0.037166   537.8   <2e-16 ***
## depth       -1.707162   0.006366  -268.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.961 on 34754 degrees of freedom
## Multiple R-squared:  0.6742, Adjusted R-squared:  0.6742
## F-statistic: 7.192e+04 on 1 and 34754 DF,  p-value: < 2.2e-16
```

```
cor.test(NTL.LTER.JUL$temperature_C, NTL.LTER.JUL$depth)
```

```
##
##  Pearson's product-moment correlation
##
## data:  NTL.LTER.JUL$temperature_C and NTL.LTER.JUL$depth
## t = -268.17, df = 34754, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8244884 -0.8176373
## sample estimates:
##       cor
## -0.8210924
```

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: Looking at the R-squared value for the regression indicates that 67% of the variability in temperature is explained by increasing depth. The degrees of freedom are 34754 indicating that the fit is better since there are a lot of independent values. The p-value is less than .05 indicating that this relationship is statistically significant. Based on the corelation coeffecient, it appears that temperature is predicted to change by -1.7 for every 1m change 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 AIC
temp.depth.AIC <- lm(data = NTL.LTER.JUL, temperature_C ~ depth + year4 +
                       daynum)
step(temp.depth.AIC)
```

```
## Start:  AIC=92515.66
## temperature_C ~ depth + year4 + daynum
##
##          Df Sum of Sq     RSS    AIC
## <none>                  497693  92516
## - year4   1       167  497861  92525
## - daynum  1     47378  545071  95674
## - depth   1   1130140 1627834 133700

##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = NTL.LTER.JUL)
##
## Coefficients:
## (Intercept)        depth        year4       daynum
##   -2.268081    -1.708651     0.007734     0.034990
```

```
#10 multiple lm
temp.depth.multi.lm <- lm(data = NTL.LTER.JUL, temperature_C ~ depth + year4 +
                            daynum)
summary(temp.depth.multi.lm)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = NTL.LTER.JUL)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.7228  -2.8606  -0.1706   2.9267  17.8338
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -2.2680808  4.5365880   -0.500 0.617111
## depth       -1.7086514  0.0060824 -280.915  < 2e-16 ***
## year4        0.0077342  0.0022622    3.419 0.000629 ***
## daynum       0.0349904  0.0006083   57.517  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.784 on 34752 degrees of freedom
## Multiple R-squared:  0.7026, Adjusted R-squared:  0.7026
## F-statistic: 2.737e+04 on 3 and 34752 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: as july increases temp does up. years temp goes up global warming, and depth is a significant impact on temperature..... r squared value of XX. the multiple variable regression is better as if covers more of the variablitity from 67% to 70%.

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 make anova then format as lm
temp.depth.avo <- aov(data = NTL.LTER.JUL, temperature_C ~ lakename)
summary(temp.depth.avo)
```

```
##                Df  Sum Sq Mean Sq F value Pr(>F)
## lakename        8   57921    7240   155.7 <2e-16 ***
## Residuals   34747 1615571      46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
temp.depth.lm <- lm(data = NTL.LTER.JUL, temperature_C ~ lakename)
summary(temp.depth.lm)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = NTL.LTER.JUL)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.436  -5.959  -2.559   6.549  24.321
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 16.7363     0.3240  51.660  < 2e-16 ***
## lakenameCrampton Lake       -2.5443     0.3833  -6.638 3.23e-11 ***
## lakenameEast Long Lake      -6.9570     0.3436 -20.248  < 2e-16 ***
## lakenameHummingbird Lake    -6.6985     0.4775 -14.030  < 2e-16 ***
## lakenamePaul Lake           -3.9441     0.3316 -11.893  < 2e-16 ***
## lakenamePeter Lake          -4.4838     0.3309 -13.549  < 2e-16 ***
## lakenameTuesday Lake        -6.3896     0.3368 -18.974  < 2e-16 ***
## lakenameWard Lake           -4.3083     0.4395  -9.802  < 2e-16 ***
## lakenameWest Long Lake      -5.6778     0.3423 -16.587  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.819 on 34747 degrees of freedom
## Multiple R-squared:  0.03461,    Adjusted R-squared:  0.03439
## F-statistic: 155.7 on 8 and 34747 DF,  p-value: < 2.2e-16
```

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

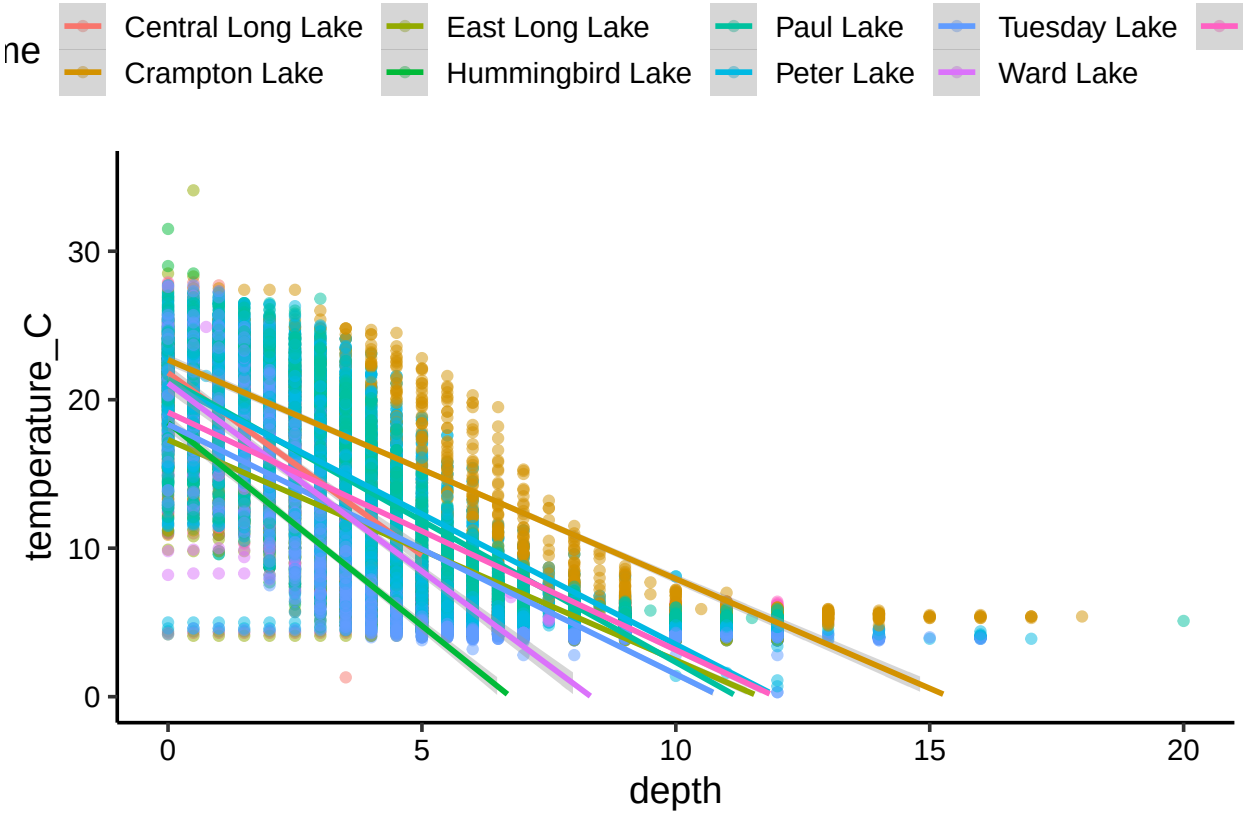
    Answer: In the anova, since the p-vale is less than .05 - we reject the null hypothesis that the temperature mean is the same across lake names. This means we would then beleive that the temperature means for each lake are different and statistically significant. The lm summary table then shows us that the temperature means are estimated with differences that are statistically signficant with another p-value less than .05.

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. graph of temp by depth separated by lake name
temp.depth.lakename <- ggplot(NTL.LTER.JUL, aes(x = depth , y = temperature_C,
                                                col = lakename)) +
  geom_point(alpha = 0.5) + geom_smooth(method = lm) + ylim(0, 35)

print(temp.depth.lakename)
```

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

```
## Warning: Removed 126 rows containing missing values (geom_smooth).
```



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

```
#15 turkey test for lake's temp means
TukeyHSD(temp.depth.avo)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
```

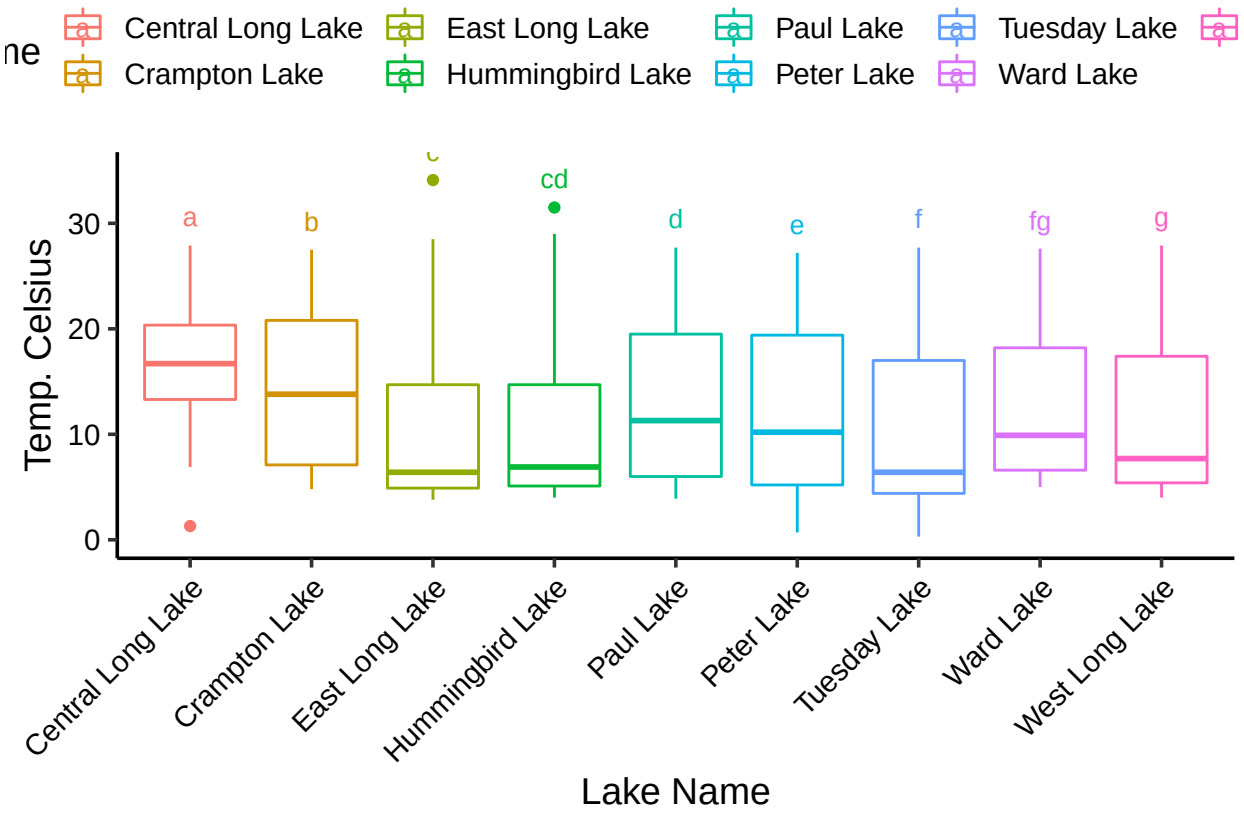
```
## Fit: aov(formula = temperature_C ~ lakename, data = NTL.LTER.JUL)
##
## $lakename
##                                          diff        lwr        upr     p adj
## Crampton Lake-Central Long Lake    -2.5442854 -3.7331780 -1.3553927 0.0000000
## East Long Lake-Central Long Lake   -6.9570473 -8.0227648 -5.8913298 0.0000000
## Hummingbird Lake-Central Long Lake -6.6985124 -8.1794348 -5.2175900 0.0000000
## Paul Lake-Central Long Lake        -3.9440682 -4.9727040 -2.9154324 0.0000000
## Peter Lake-Central Long Lake       -4.4837864 -5.5102613 -3.4573116 0.0000000
## Tuesday Lake-Central Long Lake     -6.3896413 -7.4341675 -5.3451152 0.0000000
## Ward Lake-Central Long Lake        -4.3082596 -5.6715463 -2.9449730 0.0000000
## West Long Lake-Central Long Lake   -5.6777623 -6.7395105 -4.6160141 0.0000000
## East Long Lake-Crampton Lake       -4.4127620 -5.1405823 -3.6849417 0.0000000
## Hummingbird Lake-Crampton Lake     -4.1542271 -5.4140285 -2.8944256 0.0000000
## Paul Lake-Crampton Lake            -1.3997828 -2.0721371 -0.7274286 0.0000000
## Peter Lake-Crampton Lake           -1.9395011 -2.6085446 -1.2704575 0.0000000
## Tuesday Lake-Crampton Lake         -3.8453560 -4.5417779 -3.1489341 0.0000000
## Ward Lake-Crampton Lake            -1.7639743 -2.8831342 -0.6448143 0.0000357
## West Long Lake-Crampton Lake       -3.1334769 -3.8554727 -2.4114812 0.0000000
## Hummingbird Lake-East Long Lake     0.2585349 -0.8857499  1.4028198 0.9987916
## Paul Lake-East Long Lake            3.0129792  2.5954288  3.4305296 0.0000000
## Peter Lake-East Long Lake           2.4732609  2.0610627  2.8854591 0.0000000
## Tuesday Lake-East Long Lake         0.5674060  0.1121132  1.0226989 0.0035472
## Ward Lake-East Long Lake            2.6487877  1.6614645  3.6361109 0.0000000
## West Long Lake-East Long Lake       1.2792850  0.7857610  1.7728091 0.0000000
## Paul Lake-Hummingbird Lake          2.7544443  1.6446129  3.8642756 0.0000000
## Peter Lake-Hummingbird Lake         2.2147260  1.1068972  3.3225548 0.0000000
## Tuesday Lake-Hummingbird Lake       0.3088711 -0.8157039  1.4334461 0.9952041
## Ward Lake-Hummingbird Lake          2.3902528  0.9647057  3.8157999 0.0000071
## West Long Lake-Hummingbird Lake     1.0207501 -0.1198389  2.1613391 0.1224797
## Peter Lake-Paul Lake               -0.5397183 -0.8434372 -0.2359993 0.0000013
## Tuesday Lake-Paul Lake             -2.4455731 -2.8056140 -2.0855323 0.0000000
## Ward Lake-Paul Lake                -0.3641914 -1.3113688  0.5829859 0.9582889
## West Long Lake-Paul Lake           -1.7336941 -2.1410071 -1.3263812 0.0000000
## Tuesday Lake-Peter Lake            -1.9058549 -2.2596746 -1.5520351 0.0000000
## Ward Lake-Peter Lake                0.1755268 -0.7693033  1.1203570 0.9997136
## West Long Lake-Peter Lake          -1.1939759 -1.5958003 -0.7921515 0.0000000
## Ward Lake-Tuesday Lake              2.0813817  1.1169709  3.0457925 0.0000000
## West Long Lake-Tuesday Lake         0.7118790  0.2659563  1.1578017 0.0000259
## West Long Lake-Ward Lake           -1.3695027 -2.3525401 -0.3864652 0.0005266
```

```
#make a boxplot to more easily see the tukey results
tukey.boxplot.groups <- HSD.test(temp.depth.avo, "lakename", group = TRUE)
tukey.boxplot.groups
```

```
## $statistics
##    MSerror    Df     Mean       CV
##   46.49528 34747 11.80871 57.74336
##
## $parameters
##    test   name.t ntr StudentizedRange alpha
##   Tukey lakename   9         4.386509  0.05
##
## $means
```

```
##                   temperature_C      std     r Min  Max  Q25  Q50   Q75
## Central Long Lake     16.736343 4.540842   443 1.3 27.9 13.3 16.7 20.35
## Crampton Lake         14.192058 6.801706  1108 4.8 27.5  7.1 13.8 20.80
## East Long Lake         9.779296 6.304109  3550 3.8 34.1  4.9  6.4 14.70
## Hummingbird Lake      10.037831 6.117160   378 4.0 31.5  5.1  6.9 14.70
## Paul Lake             12.792275 6.783047  9253 3.9 27.7  6.0 11.3 19.50
## Peter Lake            12.252557 7.119817 10189 0.7 27.2  5.2 10.2 19.40
## Tuesday Lake          10.346702 7.027998  5503 0.3 27.7  4.4  6.4 17.00
## Ward Lake             12.428083 6.575945   527 5.0 27.6  6.6  9.9 18.20
## West Long Lake        11.058581 6.555168  3805 4.0 27.9  5.4  7.7 17.40
##
## $comparison
## NULL
##
## $groups
##                   temperature_C groups
## Central Long Lake     16.736343      a
## Crampton Lake         14.192058      b
## Paul Lake             12.792275      c
## Ward Lake             12.428083     cd
## Peter Lake            12.252557      d
## West Long Lake        11.058581      e
## Tuesday Lake          10.346702      f
## Hummingbird Lake      10.037831     fg
## East Long Lake         9.779296      g
##
## attr(,"class")
## [1] "group"
```

```
tukey.boxplot <- ggplot(NTL.LTER.JUL, aes(x = lakename, y = temperature_C,
                                          color = lakename)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  stat_summary(geom = "text", fun = max, vjust = -1, size = 3.5,
               label = c("a", "b", "c", "cd", "d", "e",
                         "f", "fg", "g")) +
  labs(x = "Lake Name", y = "Temp. Celsius") +
  ylim(0, 35)
print(tukey.boxplot)
```



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 Late does not have the same mean temperature as another lake, however it is quite close to Paul Lake. No lake has a mean temperature that is statisitically different / distinct 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: The HSD.test can be helpful for reviewing just Peter and Paul lakes. In the HSD test the mean temperature for Paul lake is 12.79 and for Peter lake its 12.25, meaning they are very close to each other but do have distinct mean temperatures.