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

Keanu Valibia

Spring 2024

Contents

List of Figures

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)
library(agricolae)
library(lubridate)
library(here)
library(ggplot2)

getwd()
here()

ntl.lter.raw <- read.csv(here("~/R/R Projects/EDA_Spring2024/Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw
ntl.lter.raw$sampledate <- as.Date(ntl.lter.raw$sampledate,
    format = "%m/%d/%y")

# 2

myTheme <- theme_classic(base_size = 11) + theme(axis.text = element_text(color = "black"),
    axis.line = element_line(color = "#320E3B"),
    panel.background = element_rect(fill = "#EDE6E3"),
    panel.grid.major = element_line(color = "#36382E",
        linetype = "dotted"), plot.title = element_text(size = 15),
    axis.title.x = element_text(size = 13), axis.title.y = element_text(size = 13),
    legend.position = "right")

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: mu = 0 (There is not enough reason to reject the null hypothesis: mean temperatures recorded in July do not vary across all lakes) Ha: mu != 0 (There is reason to reject the null hypothesis: mean temperatures recorded in July do not vary 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

ntl.lter.wrangled <- ntl.lter.raw %>%
    filter(month(sampledate) %in% 5) %>%
    select(lakename, year4, daynum, depth, temperature_C) %>%
    drop_na()

# 5

tempByDepth <- ggplot(ntl.lter.wrangled, aes(x = depth,
    y = temperature_C)) + geom_smooth(method = "lm",
    color = "#320E3B") + labs(title = "Depth-to-Temperature Analysis",
    caption = "Source: Source: North Temperate Lakes Long-Term Ecological Research Station") +
    xlab("Depth") + ylab("Temperature (Celsius)") +
    ylim(0, 35) + geom_point()

tempByDepth
```

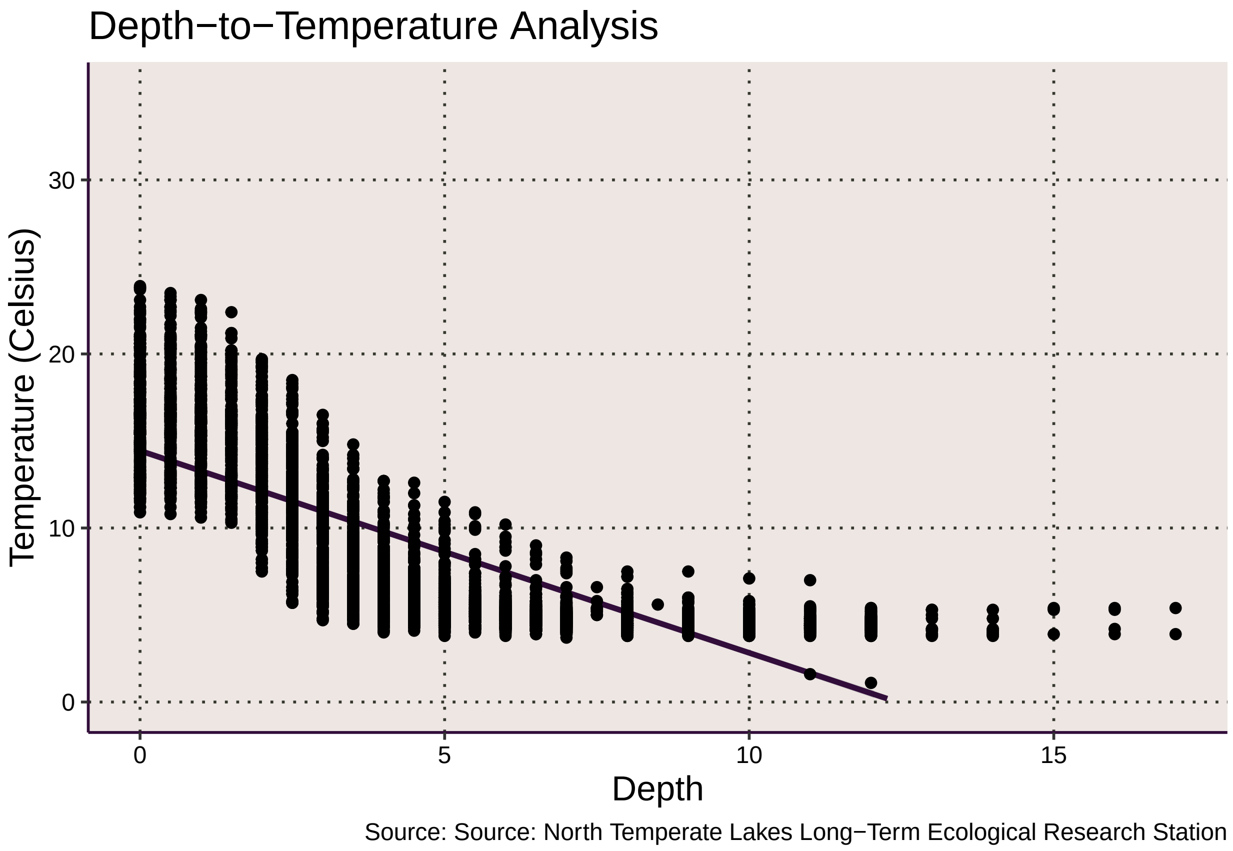


Figure 1: Depth-To-Temperature Analysis

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 scatterplot suggests a negative relationship between Temperature and Depth. However, there is a certain point in which temperatures flatten (approximately 5 degrees). Perhaps at temperatures closer to 0 (since it's freezing temperature of water), there seems to be a logical floor in which we can measure the temperature of water before it solidifies.

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

```
# 7

summary(ntl.lter.wrangled)
```

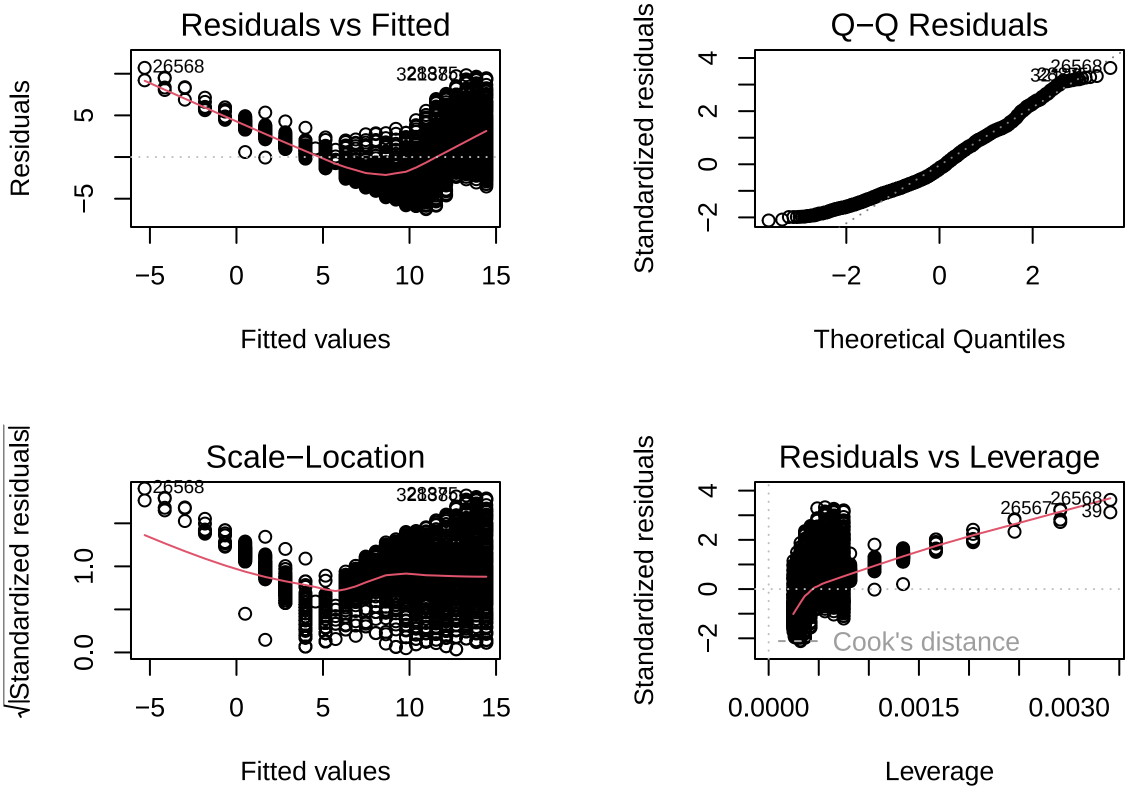
```
##    lakename             year4          daynum          depth
##  Length:4081        Min.   :1984   Min.   :124.0   Min.   : 0.000
##  Class :character   1st Qu.:1992   1st Qu.:142.0   1st Qu.: 2.000
##  Mode  :character   Median :1999   Median :146.0   Median : 4.500
##                     Mean   :2000   Mean   :144.5   Mean   : 4.847
##                     3rd Qu.:2009   3rd Qu.:148.0   3rd Qu.: 7.000
##                     Max.   :2016   Max.   :152.0   Max.   :17.000
##  temperature_C
##  Min.   : 1.100
##  1st Qu.: 4.800
##  Median : 6.300
##  Mean   : 8.809
##  3rd Qu.:12.600
##  Max.   :23.900
```

```
regression <- lm(data = subset(ntl.lter.raw, month(sampledate) %in%
    5), temperature_C ~ depth)

summary(regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = subset(ntl.lter.raw,
##     month(sampledate) %in% 5))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2541 -2.2895 -0.4476  2.0652 10.7042
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.43808    0.08093  178.41   <2e-16 ***
## depth       -1.16131    0.01369  -84.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.958 on 4079 degrees of freedom
##   (438 observations deleted due to missingness)
## Multiple R-squared:  0.6381, Adjusted R-squared:  0.638
## F-statistic:  7193 on 1 and 4079 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2), mar = c(4, 4, 4, 4))
plot(regression)
```



```r
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: R-squared is 0.638, which indicates that the model explains about 63.8% of the variability between changes in dpeth and temperature. The DF is 4079. The linear regression model results in a p-value of 2.2e-16, which is statistically significant at the 0% level. Temperature is predeicted to drop by -1.16 degrees for every 1m increase in 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.

```r
# 9

# Choose a model by AIC in a Stepwise
# Algorithm

AIC <- lm(data = ntl.lter.wrangled, temperature_C ~
    depth + year4 + daynum)
AIC
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = ntl.lter.wrangled)
##
## Coefficients:
## (Intercept)        depth        year4       daynum
##    23.31590     -1.16222     -0.01097      0.09041
```

```r
step(AIC)
```

```
## Start:  AIC=8745.05
## temperature_C ~ depth + year4 + daynum
##
##          Df Sum of Sq   RSS     AIC
## <none>                34718  8745.1
## - year4   1        45 34763  8748.3
## - daynum  1       898 35615  8847.2
## - depth   1     63008 97726 12966.5
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = ntl.lter.wrangled)
##
## Coefficients:
## (Intercept)        depth        year4       daynum
##    23.31590     -1.16222     -0.01097      0.09041
```

```r
# 10

AICmodel <- lm(data = ntl.lter.wrangled, temperature_C ~
    depth + year4 + daynum)
summary(AICmodel)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = ntl.lter.wrangled)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4799 -2.2645 -0.4029  2.0595 10.3542
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.315899   9.709021   2.401   0.0164 *
## depth       -1.162220   0.013511 -86.019   <2e-16 ***
## year4       -0.010970   0.004782  -2.294   0.0218 *
## daynum       0.090413   0.008806  10.267   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.918 on 4077 degrees of freedom
## Multiple R-squared:  0.6479, Adjusted R-squared:  0.6476
## F-statistic:  2501 on 3 and 4077 DF,  p-value: < 2.2e-16
```

```
AICmodel
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = ntl.lter.wrangled)
##
## Coefficients:
## (Intercept)        depth        year4       daynum
##    23.31590     -1.16222     -0.01097      0.09041
```

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 are the same as the original input: year, depth, and day number. The R-squared is 0.6476, which means that 64.76% of the variation is explained by the model. The p-values are the same, however the AIC model has a slight increase of the Adjusted R-Squared which indicates that the AIC model slightly explains the variability more than the linear regression model. Additionally, the coefficients for depth only decreases by ~0.001 from a linear regression model to an AIC model. These differences are only marginal.

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

lake.anova <- aov(data = ntl.lter.wrangled, temperature_C ~
    lakename)

summary(lake.anova)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## lakename       8   2066  258.29   10.89 2.61e-15 ***
## Residuals   4072  96533   23.71
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
lake.lmRegression <- lm(data = ntl.lter.wrangled,
    temperature_C ~ lakename)
summary(lake.lmRegression)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = ntl.lter.wrangled)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.725 -3.925 -2.464  3.704 15.836
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 12.9065     0.7179  17.978  < 2e-16 ***
## lakenameCrampton Lake       -3.1428     0.8406  -3.739 0.000187 ***
## lakenameEast Long Lake      -4.9425     0.7551  -6.546 6.65e-11 ***
## lakenameHummingbird Lake   -5.3675     1.0457  -5.133 2.99e-07 ***
## lakenamePaul Lake           -3.6106     0.7329  -4.926 8.71e-07 ***
## lakenamePeter Lake          -4.0813     0.7312  -5.582 2.54e-08 ***
## lakenameTuesday Lake        -4.9540     0.7451  -6.649 3.34e-11 ***
## lakenameWard Lake           -3.1084     0.8561  -3.631 0.000286 ***
## lakenameWest Long Lake      -4.1856     0.7565  -5.533 3.35e-08 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.869 on 4072 degrees of freedom
## Multiple R-squared:  0.02096,    Adjusted R-squared:  0.01903
## F-statistic:  10.9 on 8 and 4072 DF,  p-value: 2.609e-15
```

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

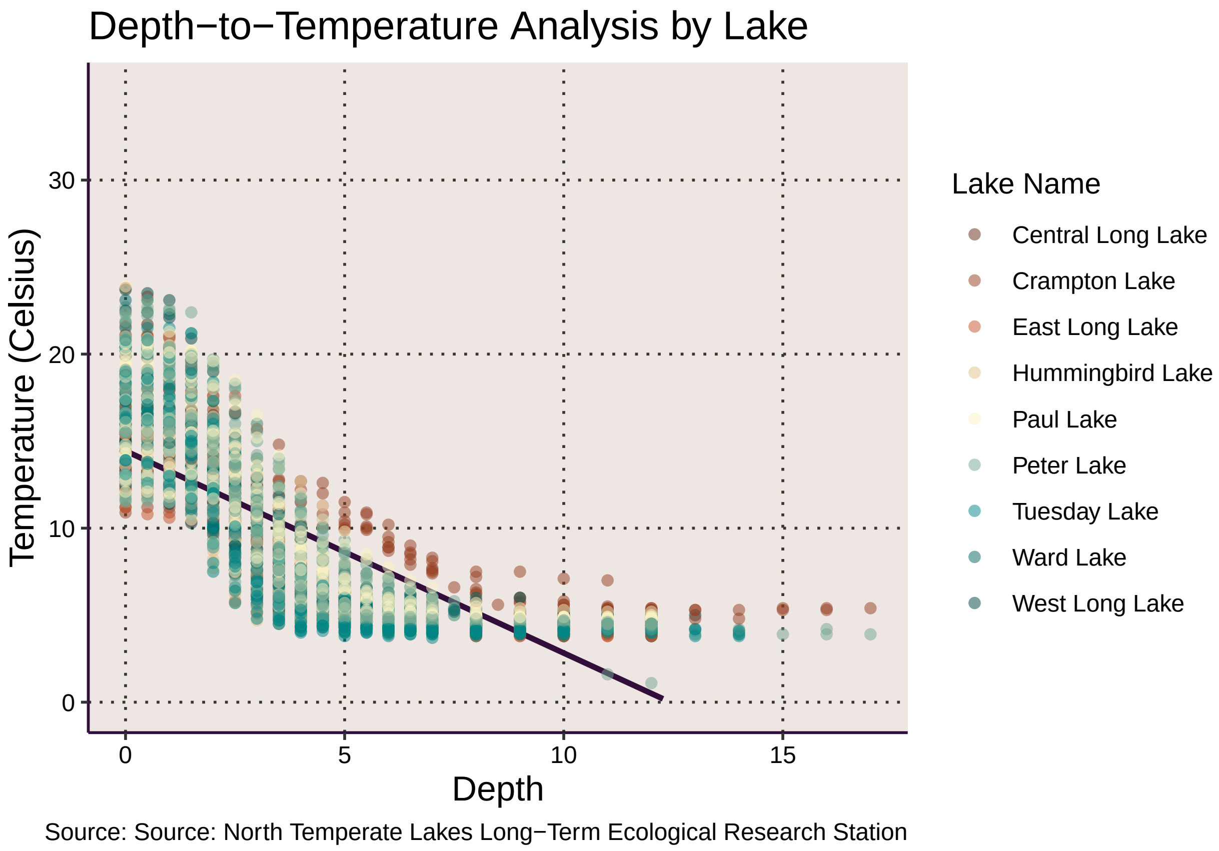
    Answer: Yes, there is a significant difference in mean temperatures among the lakes. The overall model p-value is significant at the 0% level with a p-value of 2.61e-15. When examined at the coefficient-level, all lakes have a significant difference all at the 0% level.

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.

lake.tempByDepth.plot <- ggplot(ntl.lter.wrangled,
    aes(x = depth, y = temperature_C, color = lakename)) +
    geom_smooth(method = "lm", se = FALSE, color = "#320E3B") +
    labs(title = "Depth-to-Temperature Analysis by Lake",
        caption = "Source: Source: North Temperate Lakes Long-Term Ecological Research Station",
        color = "Lake Name") + xlab("Depth") +
    ylab("Temperature (Celsius)") + ylim(0, 35) +
    geom_point(alpha = 0.5) + scale_color_manual(values = c("#642915",
    "#963e20", "#c7522a", "#e5c185", "#fbf2c4",
    "#74a892", "#008585", "#006464", "#004343"))
```

```
lake.tempByDepth.plot
```



Source: Source: North Temperate Lakes Long−Term Ecological Research Station

Figure 2: Depth-To-Temperature Analysis by Lake

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

```
# 15

lake.tukey <- TukeyHSD(lake.anova)
lake.tukey
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = ntl.lter.wrangled)
##
## $lakename
##                                          diff        lwr        upr     p adj
## Crampton Lake-Central Long Lake    -3.14281206 -5.7514206 -0.5342035 0.0058482
## East Long Lake-Central Long Lake   -4.94254945 -7.2858020 -2.5992969 0.0000000
## Hummingbird Lake-Central Long Lake -5.36749735 -8.6128582 -2.1221365 0.0000106
## Paul Lake-Central Long Lake        -3.61056214 -5.8850273 -1.3360970 0.0000307
## Peter Lake-Central Long Lake       -4.08133837 -6.3506138 -1.8120629 0.0000009
## Tuesday Lake-Central Long Lake     -4.95400496 -7.2662810 -2.6417289 0.0000000
## Ward Lake-Central Long Lake        -3.10835660 -5.7650895 -0.4516237 0.0087221
## West Long Lake-Central Long Lake   -4.18560828 -6.5334543 -1.8377622 0.0000012
## East Long Lake-Crampton Lake       -1.79973739 -3.3387671 -0.2607076 0.0087808
## Hummingbird Lake-Crampton Lake     -2.22468529 -4.9468438  0.4974732 0.2148424
## Paul Lake-Crampton Lake            -0.46775008 -1.8998714  0.9643713 0.9847007
## Peter Lake-Crampton Lake           -0.93852630 -2.3623911  0.4853385 0.5113037
## Tuesday Lake-Crampton Lake         -1.81119290 -3.3026353 -0.3197505 0.0052244
## Ward Lake-Crampton Lake             0.03445546 -1.9494824  2.0183934 1.0000000
## West Long Lake-Crampton Lake       -1.04279622 -2.5888108  0.5032183 0.4780238
## Hummingbird Lake-East Long Lake    -0.42494790 -2.8939843  2.0440885 0.9998375
## Paul Lake-East Long Lake            1.33198731  0.4735202  2.1904544 0.0000534
## Peter Lake-East Long Lake           0.86121109  0.0165898  1.7058324 0.0416184
## Tuesday Lake-East Long Lake        -0.01145551 -0.9656015  0.9426905 1.0000000
## Ward Lake-East Long Lake            1.83419285  0.2149326  3.4534531 0.0131761
## West Long Lake-East Long Lake       0.75694118 -0.2804378  1.7943201 0.3642027
## Paul Lake-Hummingbird Lake          1.75693521 -0.6469159  4.1607863 0.3618321
## Peter Lake-Hummingbird Lake         1.28615898 -1.1127823  3.6851003 0.7686653
## Tuesday Lake-Hummingbird Lake       0.41349239 -2.0261651  2.8531499 0.9998553
## Ward Lake-Hummingbird Lake          2.25914075 -0.5091689  5.0274503 0.2165387
## West Long Lake-Hummingbird Lake     1.18188907 -1.2915073  3.6552854 0.8637646
## Peter Lake-Paul Lake               -0.47077622 -1.0998582  0.1583058 0.3286138
## Tuesday Lake-Paul Lake             -1.34344282 -2.1133477 -0.5735379 0.0000023
## Ward Lake-Paul Lake                 0.50220554 -1.0158072  2.0202183 0.9833881
## West Long Lake-Paul Lake           -0.57504613 -1.4459733  0.2958811 0.5088458
## Tuesday Lake-Peter Lake            -0.87266660 -1.6271021 -0.1182311 0.0101222
## Ward Lake-Peter Lake                0.97298176 -0.5372441  2.4832076 0.5438174
## West Long Lake-Peter Lake          -0.10426991 -0.9615526  0.7530128 0.9999887
## Ward Lake-Tuesday Lake              1.84564836  0.2715481  3.4197486 0.0084773
## West Long Lake-Tuesday Lake         0.76839668 -0.1969752  1.7337686 0.2468682
## West Long Lake-Ward Lake           -1.07725168 -2.7031521  0.5486487 0.5038692
```

```
tukey.lakeGroups <- HSD.test(lake.anova, "lakename",
    group = TRUE)
tukey.lakeGroups
```

```
## $statistics
##    MSerror   Df     Mean       CV
##   23.70664 4072 8.808944 55.27276
## 
## $parameters
##    test   name.t ntr StudentizedRange alpha
##   Tukey lakename   9         4.388885  0.05
## 
## $means
##                    temperature_C      std    r        se Min  Max Q25  Q50
## Central Long Lake      12.906522 3.840567   46 0.7178870 6.9 21.0 9.6 12.5
## Crampton Lake           9.763710 3.544430  124 0.4372443 4.8 17.0 6.0 10.1
## East Long Lake          7.963972 4.737117  433 0.2339866 3.8 23.8 4.5  5.5
## Hummingbird Lake        7.539024 4.007423   41 0.7604017 4.4 15.4 4.5  5.0
## Paul Lake               9.295960 4.871536 1089 0.1475438 4.3 23.9 5.0  7.0
## Peter Lake              8.825183 4.995650 1227 0.1389993 1.1 22.7 4.6  6.3
## Tuesday Lake            7.952517 5.033933  596 0.1994398 3.7 21.8 4.2  5.0
## Ward Lake               9.798165 5.013408  109 0.4663605 5.2 23.1 5.6  7.2
## West Long Lake         8.720913 4.854219  416 0.2387197 4.3 23.7 4.9  6.0
##                       Q75
## Central Long Lake 16.175
## Crampton Lake     12.700
## East Long Lake    11.000
## Hummingbird Lake  11.700
## Paul Lake         13.100
## Peter Lake        12.750
## Tuesday Lake      12.000
## Ward Lake         13.400
## West Long Lake    12.625
##
## $comparison
## NULL
##
## $groups
##                   temperature_C groups
## Central Long Lake     12.906522      a
## Ward Lake              9.798165      b
## Crampton Lake          9.763710      b
## Paul Lake              9.295960      b
## Peter Lake             8.825183      b
## West Long Lake         8.720913     bc
## East Long Lake         7.963972      c
## Tuesday Lake           7.952517      c
## Hummingbird Lake       7.539024      c
##
## 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: The lakes that have the closest statistically significant mean temperatures at the 95% Confidence Level are Peter Lake and East Long Lake (Diff: 0.86 at p-value: 0.042). However, Ward Lake, Crampton Lake, and Paul Lake are all grouped similarly to Peter Lake. Central Lake has a mean temperature that is statistically different 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: We may use a two-sided t-test to compare the means of the two lakes, using Peter and Paul as the categorical variable.

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?

```r
ntl.lter.wrangled2 <- ntl.lter.wrangled %>%
    filter(lakename %in% c("Crampton Lake", "Ward Lake"))
```

```r
ntl.twoSample <- t.test(temperature_C ~ lakename,
    ntl.lter.wrangled2)
ntl.twoSample
```

```
##
##  Welch Two Sample t-test
##
## data:  temperature_C by lakename
## t = -0.059807, df = 191.32, p-value = 0.9524
## alternative hypothesis: true difference in means between group Crampton Lake and group Ward Lake is
## 95 percent confidence interval:
##  -1.170800  1.101889
## sample estimates:
## mean in group Crampton Lake     mean in group Ward Lake
##                    9.763710                    9.798165
```

Answer: The mean tempreatures for the lakes seem to be equal, but are not signficant at the 95% Confidence Level (p-value = 0.9524). There is not enough evidence to reject the null hypothesis (mu = 0). These results match the answer I received in part 16, which indicates a mean temperature difference of ~0.03 that is not statistically significant at the 95% Confidence Level.