

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

Reed Leon-Hinton

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 Tuesday, March 2 at 1:00 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

```
remove(list = ls()) # clearing the environment as always, it's still a pet peeve
getwd() # showing the correct working directory.
```

```
## [1] "C:/Users/shado/Documents/Graduate School Stuff/ENVIRON 872 - Environmental Data Analytics/Envir
```

```
library(tidyverse)
```

```
## -- Attaching packages ----- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3      v purrr   0.3.4
## v tibble  3.0.5      v dplyr   1.0.3
## v tidyr   1.1.2      v stringr 1.4.0
## v readr   1.4.0      v forcats 0.5.1
```

```
## -- Conflicts ----- tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

```
library(agricolae)
```

```
library(lubridate) # loading lubridate to enable easy manipulation of date objects
```

```
##
```

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

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

raw_chemistry <- read.csv(file = "../Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv",
                          stringsAsFactors = TRUE)
head(raw_chemistry) # checking to see the column names for the file

##   lakeid  lakename year4 daynum sampledte depth temperature_C dissolvedOxygen
## 1      L Paul Lake 1984   148   5/27/84  0.00             14.5             9.5
## 2      L Paul Lake 1984   148   5/27/84  0.25              NA             NA
## 3      L Paul Lake 1984   148   5/27/84  0.50              NA             NA
## 4      L Paul Lake 1984   148   5/27/84  0.75              NA             NA
## 5      L Paul Lake 1984   148   5/27/84  1.00             14.5             8.8
## 6      L Paul Lake 1984   148   5/27/84  1.50              NA             NA
##   irradianceWater irradianceDeck comments
## 1              1750             1620    <NA>
## 2              1550             1620    <NA>
## 3              1150             1620    <NA>
## 4               975             1620    <NA>
## 5               870             1620    <NA>
## 6               610             1620    <NA>

class(raw_chemistry$sampledte) # looking at how the date column imported, it was a factor

## [1] "factor"

raw_chemistry$sampledte <- mdy(raw_chemistry$sampledte) # changing the date column to a date object.
class(raw_chemistry$sampledte) # checking again to see if the change worked, it did.

## [1] "Date"

#2

# bringing back a variation on the classic "super_cool_theme"
super_cool_theme_mk2 <- theme_classic(base_size = 13) +
  theme(axis.text = element_text(color = "firebrick3"),
        legend.position = "top",
        axis.title = element_text(color = "black"),
        panel.background = element_rect(fill = "white", color = "black"),
        panel.grid.major = element_line(size = 0.5, color = "gray93"))

# setting the super_cool_theme_mk2 as the default
theme_set(super_cool_theme_mk2)
```

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: Lake temperature and depth are not correlated; slope is 0. Ha: Lake temperature and depth are correlated; slope \neq 0.
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)

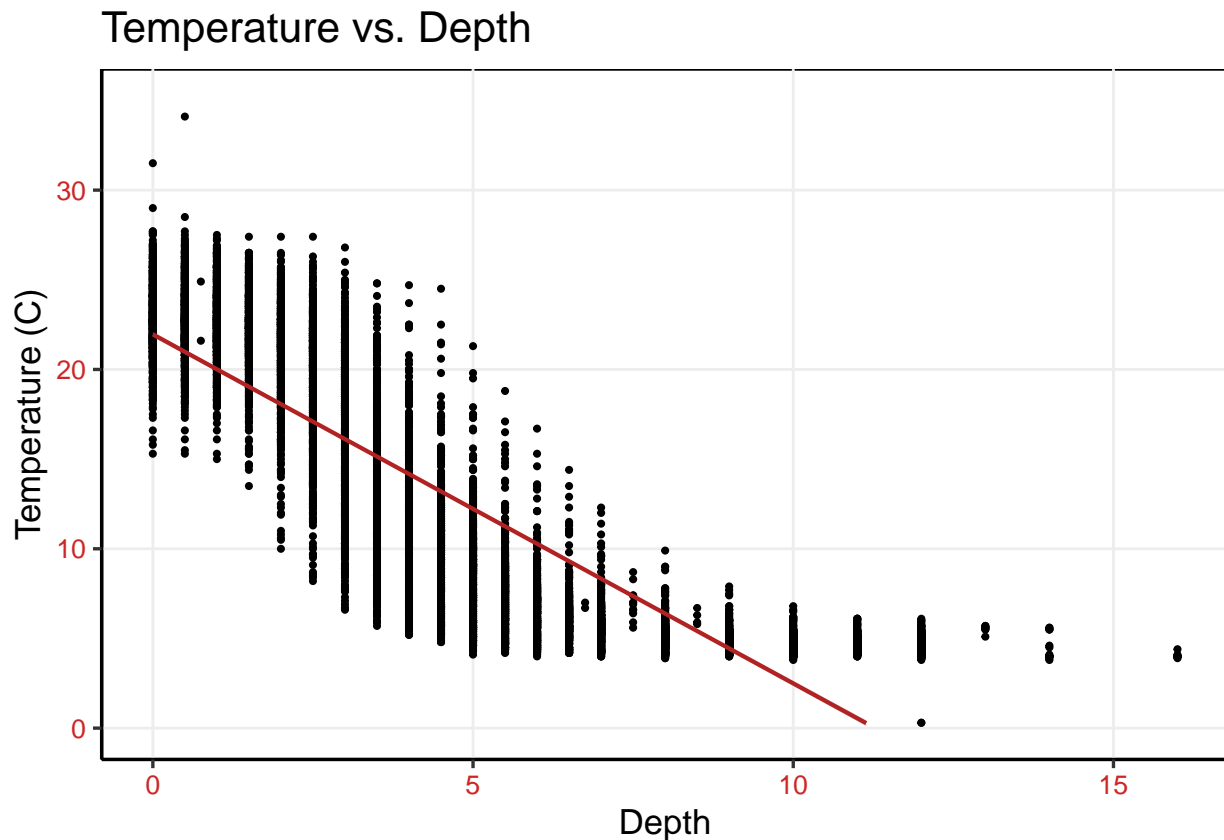
5. Visualize the relationship among the two continuous variables with a scatter plot of temperature by depth. Add a smoothed line showing the linear model, and limit temperature values from 0 to 35 °C. Make this plot look pretty and easy to read.

```
#4
raw_chemistry_filt <-
  raw_chemistry %>%
  mutate(month = month(sampledate, label = TRUE, abbr = FALSE)) %>%
  filter(month == "July") %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  drop_na()

#5
ggplot(data = raw_chemistry_filt,
       aes(x = depth, y = temperature_C)) +
  geom_point(size = .75) +
  geom_smooth(method = lm, color = "firebrick", size = .75) +
  ylim(0, 35) +
  labs(y = "Temperature (C)", x = "Depth", title = "Temperature vs. Depth")
```

```
## `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: There is a negative correlation, suggested, by this figure with regards to how depth affects

temperature. The points are distributed in a way which also suggests a non-linear relationship between depth and temperature. This suggestion is logical, as temperature will not reach 0 °C in the bottom of a relatively shallow lake when the surface is much warmer, as water serves as an excellent conductor of heat. However, due to the loss of sunlight warming the water, the drop off of temperature is severe at first but then slows down as you reach depths greater than 5.

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

```
#7

# completing the linear regression model.
tdepth_lm <- lm(data = raw_chemistry_filt, temperature_C ~ depth)
summary(tdepth_lm)

##
## Call:
## lm(formula = temperature_C ~ depth, data = raw_chemistry_filt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.5173 -3.0192  0.0633  2.9365 13.5834
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.95597    0.06792   323.3  <2e-16 ***
## depth       -1.94621    0.01174  -165.8  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.835 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

# doing a correlation test.
cor.test(raw_chemistry_filt$temperature_C, raw_chemistry_filt$depth)

##
## Pearson's product-moment correlation
##
## data:  raw_chemistry_filt$temperature_C and raw_chemistry_filt$depth
## t = -165.83, df = 9726, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8646036 -0.8542169
## sample estimates:
##          cor
## -0.8594989
```

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 p-value of the regression model is < 0.05 ($2.2e-16$), we can say with relative certainty that temperature and depth are correlated as we have rejected the null hypothesis with this model based upon 9726 degrees of freedom. The change in depth explains approximately 73.87% of the variability in temperature, with a predicted change of approximately -1.95 degrees celcius for every 1 meter deeper.

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

creating the initial mlr model to run through the step function for AIC

```
initial_mlr <- lm(data = raw_chemistry_filt, temperature_C ~ year4 + daynum + depth)
summary(initial_mlr)
```

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

using the step AIC

```
step(initial_mlr)
```

```
## 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 = raw_chemistry_filt)
##
## Coefficients:
## (Intercept)      year4      daynum      depth
```

```
##      -8.57556      0.01134      0.03978      -1.94644
```

```
#10
```

```
# the result with year included did not seem to improve the result of the model greatly
recommended_mlr <- lm(data = raw_chemistry_filt, temperature_C ~ depth + daynum)
summary(recommended_mlr)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + daynum, data = raw_chemistry_filt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.6174 -2.9809  0.0845  2.9681 13.4406
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 14.088588   0.855505   16.468  <2e-16 ***
## depth       -1.946111   0.011685  -166.541 <2e-16 ***
## daynum       0.039836   0.004318    9.225  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.818 on 9725 degrees of freedom
## Multiple R-squared:  0.741, Adjusted R-squared:  0.741
## F-statistic: 1.391e+04 on 2 and 9725 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: According to the AIC method, the year variable had the least impact followed by the daynum variable. By far, the most impactful explanatory variable was the depth. This model was run with the final set of explanatory variables as depth and daynum. The model explains 74.1 % of observed variance in temperature, a minor improvement over the 73.87% resulting from the model using only depth to explain 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
```

```
# running the anova models, as we do not need to do baseline testing
lake_temp_anova <- aov(data = raw_chemistry_filt, temperature_C ~ lakename)
summary(lake_temp_anova) # the p-value is much lower than our required 0.05.
```

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

therefore we reject the null hypothesis, the lake studied makes a statistically significant difference

```
lake_temp_anova2 <- lm(data = raw_chemistry_filt, temperature_C ~ lakename)
summary(lake_temp_anova2)
```

```
##
```

```
## Call:
```

```
## lm(formula = temperature_C ~ lakename, data = raw_chemistry_filt)
```

```
##
```

```
## 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 ***
## lakenameCrampton Lake      -2.3145     0.7699  -3.006 0.002653 **
## lakenameEast Long Lake     -7.3987     0.6918 -10.695 < 2e-16 ***
## lakenameHummingbird Lake   -6.8931     0.9429  -7.311 2.87e-13 ***
## lakenamePaul Lake         -3.8522     0.6656  -5.788 7.36e-09 ***
## lakenamePeter Lake        -4.3501     0.6645  -6.547 6.17e-11 ***
## lakenameTuesday Lake      -6.5972     0.6769  -9.746 < 2e-16 ***
## lakenameWard Lake         -3.2078     0.9429  -3.402 0.000672 ***
## lakenameWest 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
```

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

Answer: Looking at the results captured in `lake_temp_anova`, we see that the p-value is $< 2e-16$, which indicates we reject the null hypothesis. When conducting an ANOVA test such as this, the null hypothesis states that all mean values amongst the different values for `lakename` are the same. Therefore, when we reject the null hypothesis, we are saying that there is a statistically significant difference in the mean temperature among the lakes included in this analysis.

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.

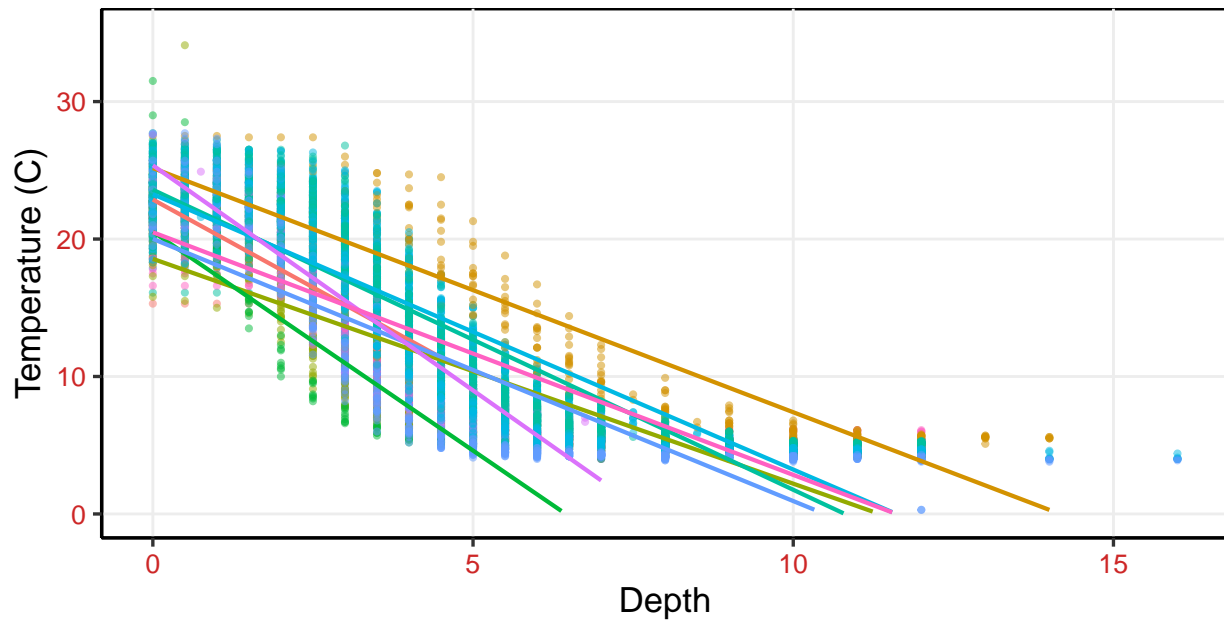
```
ggplot(data = raw_chemistry_filt,
       aes(x = depth, y = temperature_C, color = lakename)) +
  geom_point(size = .75, alpha = 0.50) +
  geom_smooth(method = lm, size = .75, se = FALSE) +
  ylim(0, 35) +
  labs(y = "Temperature (C)", x = "Depth", title = "Comparison of Lake Temperatures") +
  theme(legend.title = element_blank())
```

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

```
## Warning: Removed 73 rows containing missing values (geom_smooth).
```

Comparison of Lake Temperatures

Central Long Lake East Long Lake Paul Lake Tuesday Lake West Lon
Crampton Lake Hummingbird Lake Peter Lake Ward Lake



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

#15

```
TukeyHSD(lake_temp_anova)
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = raw_chemistry_filt)
##
## $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
```

```
lakeHSD <- HSD.test(lake_temp_anova, "lakename", group = TRUE)
lakeHSD
```

```
## $statistics
##      MSerror  Df      Mean      CV
##    54.1016 9719 12.72087 57.82135
##
## $parameters
##      test  name.t ntr StudentizedRange alpha
##    Tukey lakename   9         4.387504  0.05
##
## $means
##               temperature_C      std      r Min  Max   Q25   Q50   Q75
## Central Long Lake      17.66641 4.196292  128 8.9 26.8 14.400 18.40 21.000
## Crampton Lake          15.35189 7.244773  318 5.0 27.5  7.525 16.90 22.300
## East Long Lake         10.26767 6.766804  968 4.2 34.1  4.975  6.50 15.925
## Hummingbird Lake       10.77328 7.017845  116 4.0 31.5  5.200  7.00 15.625
## Paul Lake              13.81426 7.296928 2660 4.7 27.7  6.500 12.40 21.400
## Peter Lake             13.31626 7.669758 2872 4.0 27.0  5.600 11.40 21.500
## Tuesday Lake           11.06923 7.698687 1524 0.3 27.7  4.400  6.80 19.400
## Ward Lake              14.45862 7.409079  116 5.7 27.6  7.200 12.55 23.200
## West Long Lake         11.57865 6.980789 1026 4.0 25.7  5.400  8.00 18.800
##
## $comparison
## NULL
##
## $groups
##               temperature_C groups
## Central Long Lake      17.66641      a
## Crampton Lake          15.35189     ab
## Ward Lake              14.45862     bc
## Paul Lake              13.81426      c
```

```
## Peter Lake          13.31626      c
## West Long Lake      11.57865      d
## Tuesday Lake        11.06923     de
## Hummingbird Lake    10.77328     de
## East Long Lake      10.26767      e
##
## 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: Statistically speaking, the only lakes which have the same mean temperature as Peter Lake are Paul Lake and Ward Lake as they are close to one another with the calculated mean and also the p-value is greater than 0.05. No, there is no lake which has a statistically distinct mean when compared with all other lakes. Every lake has at least one lake where the mean could be statistically the same.

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: A quality method for looking at two distinct populations, such as Peter and Paul Lake temperature measurements at a variety of depths, is the two sample t-test. ANOVA is generally used, as stated in the lesson, for categorical explanatory variables with more than two categories. Therefore, with only Peter and Paul Lakes under analysis, moving to the t-test approach would work much better for comparing mean temperatures.