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

Logan Dye

OVERVIEW

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

Directions

1. Rename this file `<FirstLast>_A06_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
getwd()
```

```
## [1] "/home/guest/EDA-Fall2022"
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.1
## v readr   2.1.2      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
```

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

RawLakes$sampledate <- as.Date(RawLakes$sampledate, format = "%m/%d/%y")
# 2
mytheme <- theme_classic(base_size = 12) + theme(axis.text = element_text(color = "blue"),
    legend.position = "bottom")
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 statistical difference between mean lake temeratures recorded in July as the depth changes. Ha: There is a statistical difference in the mean lake temperatures recorded in July as the depth changes.
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)

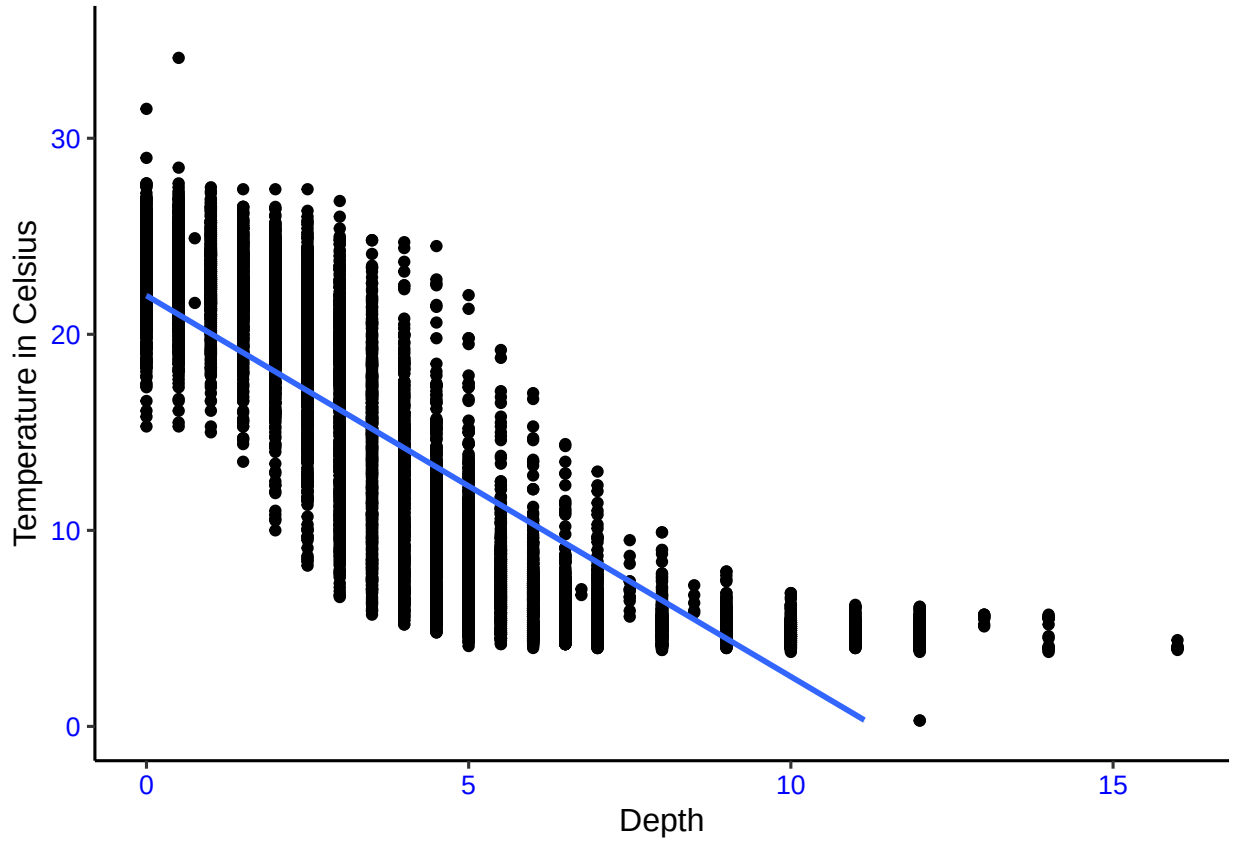
5. Visualize the relationship among the two continuous variables with a scatter plot of temperature by depth. Add a smoothed line showing the linear model, and limit temperature values from 0 to 35 °C. Make this plot look pretty and easy to read.

```
# 4
FilteredLakes <- RawLakes %>%
    filter(daynum %in% c(182:214)) %>%
    select(lakename:daynum, daynum:temperature_C) %>%
    na.omit()


# 5
LakePlot1 <- ggplot(FilteredLakes, aes(x = depth, y = temperature_C)) + geom_point() +
    geom_smooth(method = "lm") + ylim(0, 35) + xlab("Depth") + ylab("Temperature in Celsius")
print(LakePlot1)
```

```
## `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: This figure suggests a negative cooralation between deopth and Temperarature. As the depth increases the temperatures will get lower and lower. The distribution of points suggests a trend that is not totally liniear. The point structure appears to be more logarithmic.

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

```
# 7
TempRegression <- lm(data = FilteredLakes, temperature_C ~ depth)
summary(TempRegression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = FilteredLakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5448 -3.0292  0.0959  2.9677 13.5236
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.97598    0.06647   330.6   <2e-16 ***
## depth       -1.94372    0.01146  -169.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3.853 on 10255 degrees of freedom
## Multiple R-squared:  0.7371, Adjusted R-squared:  0.737
## F-statistic: 2.875e+04 on 1 and 10255 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: Our model is based on 10255 degrees of freedom. Looking at the results of the linear regression, we see that temperature is negatively correlated to depth. This relationship is significant with a p-value of only 2.2e-16. 73.7% of the variability in temperatue is explained by the changes in depth. For every 1m change in depth we predict the temperature to change -1.94 degres Celsius.

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
TempAIC <- lm(data = FilteredLakes, temperature_C ~ year4 + daynum + depth)
step(TempAIC)
```

```
## Start:  AIC=27550.19
## temperature_C ~ year4 + daynum + depth
##
##          Df Sum of Sq    RSS   AIC
## <none>                 150381 27550
## - year4   1       208 150588 27562
## - daynum  1      1664 152045 27661
## - depth   1    427130 577511 41350
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = FilteredLakes)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##   -18.23428      0.01589      0.04268     -1.94462
```

```
# 10
MultTempReg <- lm(data = FilteredLakes, temperature_C ~ year4 + daynum + depth)
summary(MultTempReg)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = FilteredLakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6759 -3.0217  0.0915  2.9952 13.6742
##
## Coefficients:
```

```
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -18.234277   8.475972   -2.151 0.031477 *
## year4         0.015888   0.004222    3.764 0.000168 ***
## daynum        0.042676   0.004006   10.652  < 2e-16 ***
## depth        -1.944619   0.011395 -170.651  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.83 on 10253 degrees of freedom
## Multiple R-squared:  0.7403, Adjusted R-squared:  0.7402
## F-statistic:  9742 on 3 and 10253 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: The final set of explanatory variables the AIC method suggests we use is year4, daynum, and depth to predict temperature. The multiple regression analysis using all three of the variables explains 74% of the variability in temperature. There is not much of an giving using only depth explained 73.7% of the variance.

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
TempANOVA1 <- aov(data = FilteredLakes, temperature_C ~ lakename)
summary(TempANOVA1)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## lakename        8  23163  2895.4   53.38 <2e-16 ***
## Residuals   10248 555884    54.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TempANOVA2 <- lm(data = FilteredLakes, temperature_C ~ lakename)
summary(TempANOVA2)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = FilteredLakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.784  -6.616  -2.684   7.667  23.852
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 17.6664     0.6510  27.138  < 2e-16 ***
## lakenameCrampton Lake       -2.1851     0.7568  -2.887 0.003896 **
## lakenameEast Long Lake      -7.4185     0.6903 -10.747  < 2e-16 ***
## lakenameHummingbird Lake    -6.5875     0.9299  -7.084 1.49e-12 ***
## lakenamePaul Lake           -3.8206     0.6661  -5.736 9.97e-09 ***
## lakenamePeter Lake          -4.3329     0.6646  -6.520 7.38e-11 ***
```

```
## lakenameTuesday Lake      -6.5823     0.6763  -9.733  < 2e-16 ***
## lakenameWard Lake         -3.2078     0.9441  -3.398 0.000682 ***
## lakenameWest Long Lake    -6.0507     0.6871  -8.806  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.365 on 10248 degrees of freedom
## Multiple R-squared:   0.04,  Adjusted R-squared:  0.03925
## F-statistic: 53.38 on 8 and 10248 DF,  p-value: < 2.2e-16
```

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

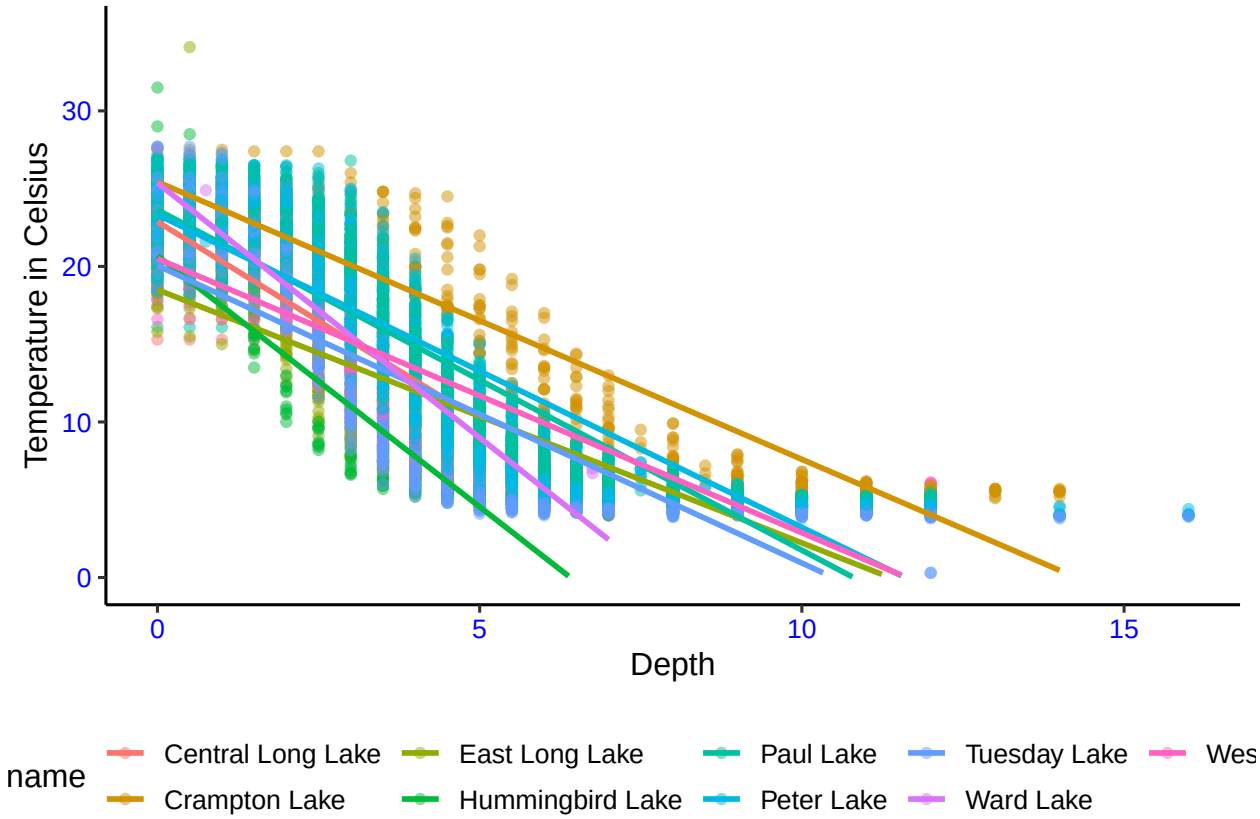
    Answer: Yes there is a significant difference in mean temperature between the lakes. The ANOVA test showed a p-value of 2e-16 in the anova model and 2.2e-16 in the linear model. Both of these p-values would indicate significant difference.

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.
LakePlot2 <- ggplot(FilteredLakes, aes(x = depth, y = temperature_C, color = lakename)) +
    geom_point(alpha = 0.5) + geom_smooth(method = "lm", se = FALSE) + ylim(0, 35) +
    xlab("Depth") + ylab("Temperature in Celsius")
print(LakePlot2)
```

```
## `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(TempANOVA1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = FilteredLakes)
##
## $lakename
##                                          diff         lwr         upr
## Crampton Lake-Central Long Lake    -2.185087569 -4.53308083  0.16290569
## East Long Lake-Central Long Lake   -7.418546328 -9.56019065 -5.27690201
## Hummingbird Lake-Central Long Lake -6.587544461 -9.47256754 -3.70252138
## Paul Lake-Central Long Lake        -3.820594597 -5.88701337 -1.75417582
## Peter Lake-Central Long Lake       -4.332907900 -6.39472309 -2.27109271
## Tuesday Lake-Central Long Lake     -6.582309715 -8.68036800 -4.48425143
## Ward Lake-Central Long Lake        -3.207785560 -6.13686334 -0.27870778
## West Long Lake-Central Long Lake   -6.050733944 -8.18232438 -3.91914351
## East Long Lake-Crampton Lake       -5.233458759 -6.62707224 -3.83984528
## Hummingbird Lake-Crampton Lake     -4.402456893 -6.78549619 -2.01941759
## Paul Lake-Crampton Lake            -1.635507029 -2.91049906 -0.36051499
## Peter Lake-Crampton Lake           -2.147820331 -3.41533760 -0.88030306
## Tuesday Lake-Crampton Lake         -4.397222147 -5.72287923 -3.07156506
## Ward Lake-Crampton Lake            -1.022697992 -3.45888657  1.41349059
## West Long Lake-Crampton Lake       -3.865646375 -5.24375955 -2.48753320
## Hummingbird Lake-East Long Lake     0.831001866 -1.34900831  3.01101204
## Paul Lake-East Long Lake            3.597951730  2.76178623  4.43411723
## Peter Lake-East Long Lake           3.085638428  2.26091540  3.91036146
## Tuesday Lake-East Long Lake         0.836236612 -0.07531961  1.74779284
## Ward Lake-East Long Lake            4.210760767  1.97277443  6.44874711
## West Long Lake-East Long Lake       1.367812384  0.38152440  2.35410037
## Paul Lake-Hummingbird Lake          2.766949864  0.66079448  4.87310525
## Peter Lake-Hummingbird Lake         2.254636561  0.15299772  4.35627540
## Tuesday Lake-Hummingbird Lake       0.005234746 -2.13197196  2.14244145
## Ward Lake-Hummingbird Lake          3.379758901  0.42251346  6.33700434
## West Long Lake-Hummingbird Lake     0.536810518 -1.63332352  2.70694455
## Peter Lake-Paul Lake               -0.512313303 -1.11531763  0.09069102
## Tuesday Lake-Paul Lake             -2.761715118 -3.47891864 -2.04451159
## Ward Lake-Paul Lake                 0.612809037 -1.55330015  2.77891822
## West Long Lake-Paul Lake           -2.230139346 -3.04020732 -1.42007137
## Tuesday Lake-Peter Lake            -2.249401815 -2.95323150 -1.54557213
## Ward Lake-Peter Lake                1.125122340 -1.03659557  3.28684025
## West Long Lake-Peter Lake          -1.717826044 -2.51607755 -0.91957454
## Ward Lake-Tuesday Lake              3.374524155  1.17821110  5.57083721
## West Long Lake-Tuesday Lake         0.531575772 -0.35610218  1.41925372
## West Long Lake-Ward Lake           -2.842948383 -5.07131555 -0.61458121
##                                        p adj
## Crampton Lake-Central Long Lake    0.0918485
## East Long Lake-Central Long Lake   0.0000000
## Hummingbird Lake-Central Long Lake 0.0000000
## Paul Lake-Central Long Lake        0.0000004
## Peter Lake-Central Long Lake       0.0000000
## Tuesday Lake-Central Long Lake     0.0000000
```

```
## Ward Lake-Central Long Lake        0.0196281
## West Long Lake-Central Long Lake   0.0000000
## East Long Lake-Crampton Lake       0.0000000
## Hummingbird Lake-Crampton Lake     0.0000004
## Paul Lake-Crampton Lake            0.0022629
## Peter Lake-Crampton Lake           0.0000053
## Tuesday Lake-Crampton Lake         0.0000000
## Ward Lake-Crampton Lake            0.9309721
## West Long Lake-Crampton Lake       0.0000000
## Hummingbird Lake-East Long Lake    0.9602971
## Paul Lake-East Long Lake           0.0000000
## Peter Lake-East Long Lake          0.0000000
## Tuesday Lake-East Long Lake        0.1023985
## Ward Lake-East Long Lake           0.0000002
## West Long Lake-East Long Lake      0.0005769
## Paul Lake-Hummingbird Lake         0.0015248
## Peter Lake-Hummingbird Lake        0.0246994
## Tuesday Lake-Hummingbird Lake      1.0000000
## Ward Lake-Hummingbird Lake         0.0117765
## West Long Lake-Hummingbird Lake    0.9976842
## Peter Lake-Paul Lake               0.1718766
## Tuesday Lake-Paul Lake             0.0000000
## Ward Lake-Paul Lake                0.9941147
## West Long Lake-Paul Lake           0.0000000
## Tuesday Lake-Peter Lake            0.0000000
## Ward Lake-Peter Lake               0.7970415
## West Long Lake-Peter Lake          0.0000000
## Ward Lake-Tuesday Lake             0.0000664
## West Long Lake-Tuesday Lake        0.6430358
## West Long Lake-Ward Lake           0.0024692
```

```
TempGroups <- HSD.test(TempANOVA1, "lakename", group = TRUE)
TempGroups
```

```
## $statistics
##    MSerror    Df     Mean       CV
##   54.24316 10248 12.73406 57.83699
##
## $parameters
##    test   name.t ntr StudentizedRange alpha
##   Tukey lakename   9         4.387453  0.05
##
## $means
##                   temperature_C      std    r Min  Max   Q25   Q50  Q75
## Central Long Lake      17.66641 4.196292  128 8.9 26.8 14.40 18.40 21.0
## Crampton Lake          15.48132 7.347999  364 5.0 27.5  7.50 17.05 22.4
## East Long Lake         10.24786 6.737382 1028 4.2 34.1  5.00  6.50 16.0
## Hummingbird Lake       11.07886 7.055590  123 4.0 31.5  5.25  7.70 16.4
## Paul Lake              13.84581 7.308747 2729 4.7 27.7  6.50 12.40 21.4
## Peter Lake             13.33350 7.693111 3030 4.0 27.0  5.60 11.40 21.5
## Tuesday Lake           11.08410 7.699962 1616 0.3 27.7  4.40  6.80 19.5
## Ward Lake              14.45862 7.409079  116 5.7 27.6  7.20 12.55 23.2
## West Long Lake         11.61567 6.956682 1123 4.0 25.7  5.40  8.10 18.8
##
## $comparison
```

```
## NULL
##
## $groups
##                    temperature_C groups
## Central Long Lake      17.66641      a
## Crampton Lake          15.48132     ab
## Ward Lake              14.45862     bc
## Paul Lake              13.84581      c
## Peter Lake             13.33350      c
## West Long Lake         11.61567      d
## Tuesday Lake           11.08410     de
## Hummingbird Lake       11.07886     de
## East Long Lake         10.24786      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: According to our Tukey HSD test the only two lakes that have the same mean temperature as Peter Lake are Paul Lake and Ward Lake. There are no lakes that have a mean temperature satistically distint from all 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: If we were just looking at Peter and Paul Lake we could have run a two sample t-test. This would have tested to see if the mean of the two lakes were equivalent or not.

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?

```
WrangledJulyData <- filter(FilteredLakes, lakename %in% c("Crampton Lake", "Ward Lake"))

Lake.twosampleT <- t.test(WrangledJulyData$temperature_C ~ WrangledJulyData$lakename)
Lake.twosampleT
```

```
##
##  Welch Two Sample t-test
##
## data:  WrangledJulyData$temperature_C by WrangledJulyData$lakename
## t = 1.2972, df = 192.4, p-value = 0.1961
## alternative hypothesis: true difference in means between group Crampton Lake and group Ward Lake is n
## 95 percent confidence interval:
##  -0.5323014  2.5776973
## sample estimates:
## mean in group Crampton Lake     mean in group Ward Lake
##                    15.48132                    14.45862
```

Answer: The test gives a p-value of 0.1961. This tells us that we cannot reject the null hyphothosis. There is no statistical difference between the two lakes in July. This matches our answer from part 16 that has both of those lakes in the same group, b.