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

Student Name

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/R/EDA Fall"
```

```
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(ggplot2)
library(lubridate)
```

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

```
# loading all necessary packages
NTL_LTR <- read.csv("./Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv")
class(NTL_LTR$sampledate)
```

```
## [1] "character"
```

```
NTL_LTR$sampledate <- as.Date(NTL_LTR$sampledate, format = "%m/%d/%y")
class(NTL_LTR$sampledate)
```

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

```
# checking the class of the data, then reformating it to date and rechecking 2
GLM_plot_theme <- theme_gray(base_size = 14) + theme(axis.text = element_text(color = "Gray"),
    legend.position = "right")
theme_set(GLM_plot_theme)
# setting plot theme
```

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: Temperature across the lakes will vary regardless of depth. There will not be a significant relationship between the variables and slope should be 0 Ha: Temperature across the lakes will decrease as depth lowers and will increase at shallower depths.
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.

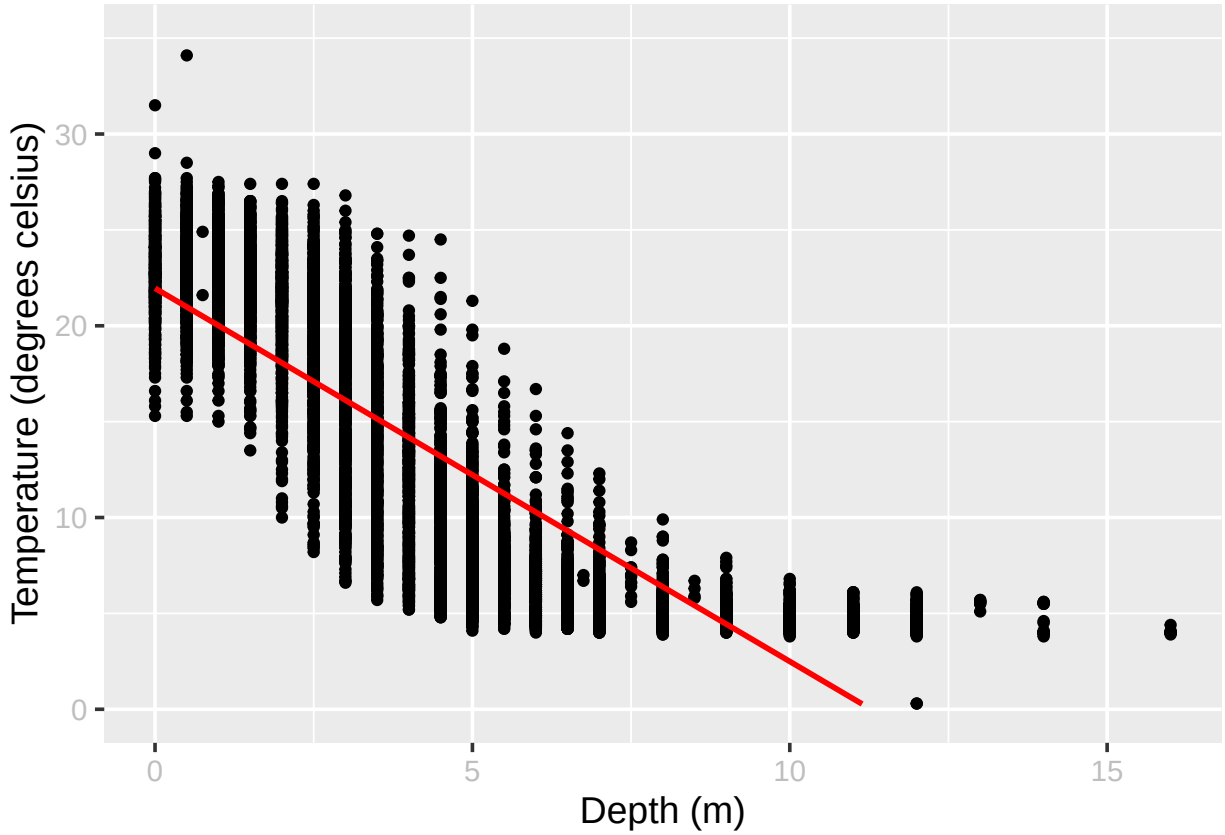
```
# 4
NTL_LTR_wrangled <- NTL_LTR %>%
    mutate(Month = month(sampledate)) %>%
    filter(Month == "7") %>%
    select("lakename", "year4", "daynum", "depth", "temperature_C") %>%
    drop_na("temperature_C")
# wrangling data to only have the selected columns and filtering out the data
# in those columns so it only has values from July, and dropping the NAs 5
NTL_LTR_Plot1 <- ggplot(NTL_LTR_wrangled, aes(x = depth, y = temperature_C)) + geom_point(aes(Color = "
    geom_smooth(method = lm, se = FALSE, color = "red") + ylim(0, 35) + GLM_plot_theme +
    xlab("Depth (m)") + ylab("Temperature (degrees celsius)")
```

```
## Warning: Ignoring unknown aesthetics: Color
```

```
print(NTL_LTR_Plot1)
```

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

```
## Warning: Removed 24 rows containing missing values (geom_smooth).
```




```
# making the first scatter plot with a single line of best fit
```

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 indicates that in general, as the depth of the lake increases, the temperature at that depth decreases in conjunction. Yes, there is a linear relationship but it is not a strong correlation at higher depths the points seem to be more varied indicating a wider range in temperatures, as the depth lowers the range of temperature decreases.

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

```
# 7

NTL_LTR_regression <- lm(NTL_LTR_wrangled$temperature_C ~ NTL_LTR_wrangled$depth)
summary(NTL_LTR_regression)
```

```
##
## Call:
## lm(formula = NTL_LTR_wrangled$temperature_C ~ NTL_LTR_wrangled$depth)
##
## 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 ***
```

```
## NTL_LTR_wrangled$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
```

```
# running a linear regression model
```

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: The model indicates about 74% of the variability in temperature is related to changes in depth and that there are 9726 degrees of freedom, that results are statistically significant since the P value is < 2.2eˆ-16. The slope is -1.946 meaning that the temperature will decrease by about 2 degrees for every one meter.

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

```
##
## Call:
## lm(formula = temperature_C ~ depth + daynum + year4, data = NTL_LTR_wrangled)
##
## 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
## depth       -1.946437   0.011683 -166.611  < 2e-16 ***
## daynum       0.039780   0.004317    9.215  < 2e-16 ***
## year4        0.011345   0.004299    2.639  0.00833 **
## ---
## 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
```

```
# creating an AIC model
step(NTL_LTR_AIC)
```

```
## Start:  AIC=26065.53
## temperature_C ~ depth + daynum + year4
##
##          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 ~ depth + daynum + year4, data = NTL_LTR_wrangled)
##
## Coefficients:
## (Intercept)        depth       daynum        year4
##    -8.57556     -1.94644      0.03978      0.01134
```

```
# using step function to see which variables should be included

# 10

NTL_Regression_multi <- lm(data = subset(NTL_LTR_wrangled), depth ~ daynum + year4)
summary(NTL_Regression_multi)
```

```
##
## Call:
## lm(formula = depth ~ daynum + year4, data = subset(NTL_LTR_wrangled))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.816 -2.734 -0.295  2.208 11.318
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.9530162  7.4913761  -0.394    0.693
## daynum      -0.0003614  0.0037471  -0.096    0.923
## year4        0.0038870  0.0037313   1.042    0.298
##
## Residual standard error: 3.313 on 9725 degrees of freedom
## Multiple R-squared:  0.0001124,  Adjusted R-squared:  -9.32e-05
## F-statistic: 0.5468 on 2 and 9725 DF,  p-value: 0.5788
```

```
# runnign a multiple regression model
```

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 is depth, daynum, and year4 The model explains 74.12% of the observed variance, this is an improvement of the previous 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
NTL_anova <- aov(data = NTL_LTR_wrangled, temperature_C ~ lakename)
summary(NTL_anova)
```

```
##               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
```

```
# running an anova model
NTL_LM <- lm(data = NTL_LTR_wrangled, temperature_C ~ lakename)
summary(NTL_LM)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = NTL_LTR_wrangled)
##
## 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
```

```
# running an lm model
```

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

    Answer: Yes there is a significant difference, since the P value is less than 2eˆ-16. The observed variance explained by the model is about 4%

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.

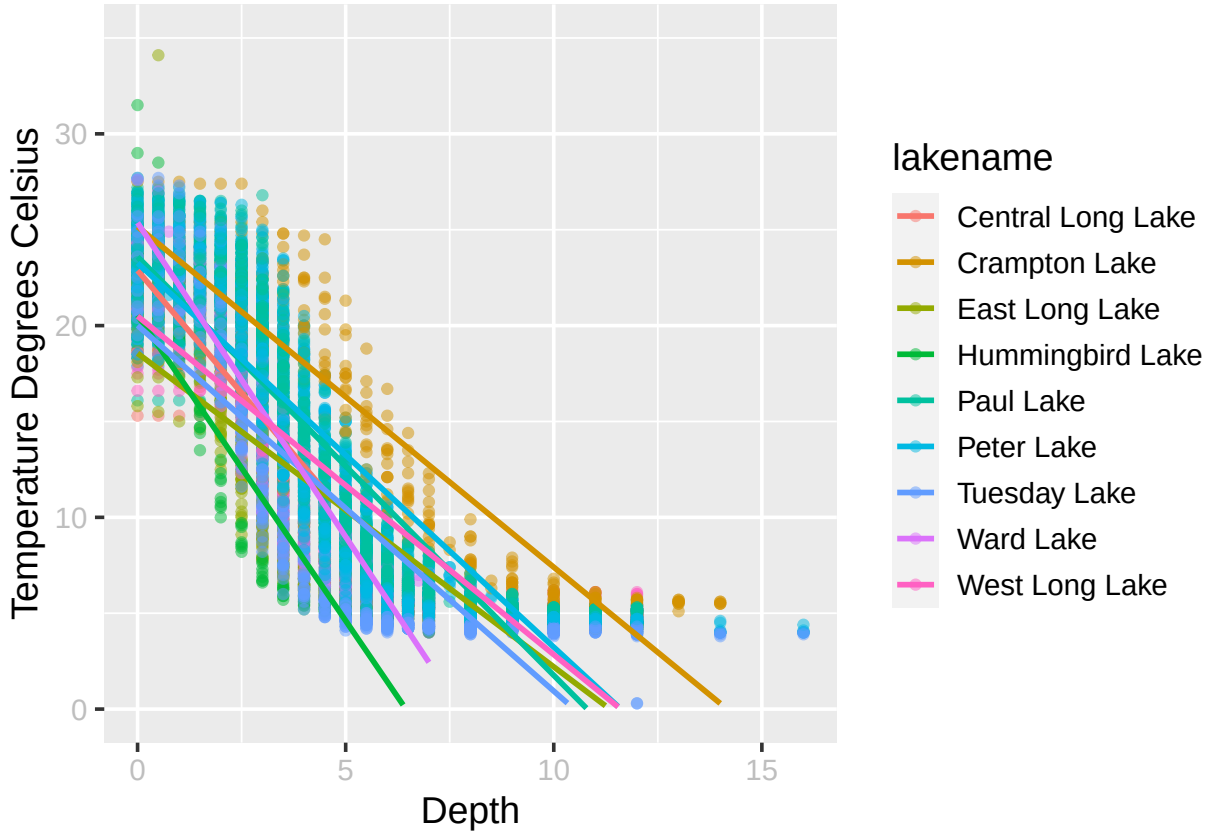
```
# 14.

NTL_LTR_plot2 <- ggplot(NTL_LTR_wrangled, aes(y = temperature_C, x = depth)) + geom_point(aes(color = la
    alpha = 0.5) + ylim(0, 35) + GLM_plot_theme + geom_smooth(method = lm, se = FALSE,
```

```
    aes(color = lakename)) + xlab("Depth") + ylab("Temperature Degrees Celsius")

print(NTL_LTR_plot2)
```

```
## `geom_smooth()` using formula 'y ~ x'
## Warning: Removed 73 rows containing missing values (geom_smooth).
```




```
# creating the second plot with variables differentiated by lakename and with
# multiple lines of regression
```

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

```
# 15
TukeyHSD(NTL_anova)
```

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

```
NTL_LTR_HSD <- HSD.test(NTL_anova, "lakename", group = TRUE)
NTL_LTR_HSD
```

```
## $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"
```

```
# running the tukey hsd test
```

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 and Ward lakes have a P value of .78 and Peter and Paul lakes have a P value of .224, meaning they are statistically similar. No lake had a mean temperature that was statistically distinct 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: Since we would only have two variables for Peter and Paul lake we could do a two sample T test .

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?

```
NTL_LTR_finalwrangle <- NTL_LTR_wrangled %>%
    filter(lakename == "Crampton Lake" | lakename == "Ward Lake") %>%
    select("lakename", "year4", "daynum", "depth", "temperature_C")
# wrangling the data to just be for Crampton and Ward lake
t.test(NTL_LTR_finalwrangle$temperature_C ~ NTL_LTR_finalwrangle$lakename, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  NTL_LTR_finalwrangle$temperature_C by NTL_LTR_finalwrangle$lakename
## t = 1.1181, df = 200.37, p-value = 0.2649
## alternative hypothesis: true difference in means between group Crampton Lake and group Ward Lake is n
## 95 percent confidence interval:
##  -0.6821129  2.4686451
## sample estimates:
## mean in group Crampton Lake     mean in group Ward Lake
##                    15.35189                    14.45862
```

```
# running the two sided t.test
```

Answer: The mean temperatures for the lake are not equal, the test says the P value between

Crampton and Ward is .2649 This does not match what we found in question 16 since the Tukey test provided us with an adjusted P value.