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

Samantha Pace

Fall 2023

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
getwd() #checking that the EDE_Fall2023 Folder is my wd
```

```
## [1] "/home/guest/R/EDE_Fall2023"
```

```
# loading packages
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(agricolae)
library(lubridate)

# importing data
NTL.LTER.raw <-
  read.csv("./Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv",
```

```
            stringsAsFactors = TRUE)

# fixing dates
NTL.LTER.raw$sampledate <- mdy(NTL.LTER.raw$sampledate)

#2
# creating my theme based on the black and white theme
mytheme <-
  theme_bw(base_size = 14)+
    theme(panel.grid.major = element_line(colour = "grey80"))+
  theme(plot.title = element_text(size = rel(1)), legend.position = "top")

# setting my theme to be the default
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: The mean lake temperature recorded during July does not change with depth across all lakes. Ha: THe mean lake temperature recorded during 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)

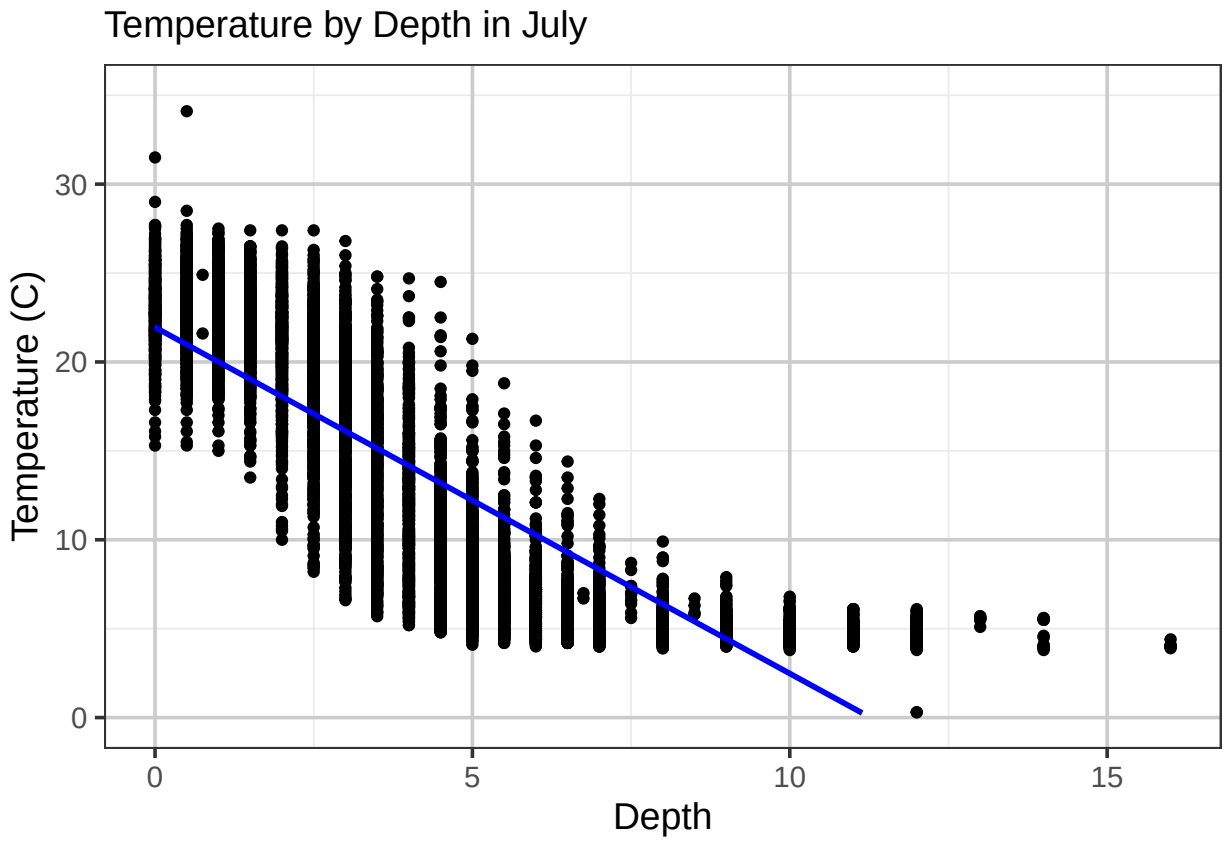
5. Visualize the relationship among the two continuous variables with a scatter plot of temperature by depth. Add a smoothed line showing the linear model, and limit temperature values from 0 to 35 °C. Make this plot look pretty and easy to read.

```
#4

#creating a new processed data set that is filtered to be only July, have certain columns,
# and be complete cases.
NTL.LTER.processed <-
  NTL.LTER.raw %>%
  filter(daynum > 181 & daynum < 213) %>% #filtering so it is only July
  select(lakename, year4, daynum, depth, temperature_C) %>% # selecting certain columns
  na.omit #dropping all incomplete cases

#5
#scatterplot of x=depth, y = temp limited to 35 degrees
#added titles, updated labels, and smoothed line of linear model
Depth.by.temp <-
  ggplot(NTL.LTER.processed, aes(x = depth, y = temperature_C))+
  geom_point()+
  ylim(0,35)+
  xlab("Depth")+
  ylab("Temperature (C)")+
  geom_smooth(method=lm, color = 'blue')+
  ggtitle("Temperature by Depth in July")
print(Depth.by.temp)
```

```
## `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: As the depth increases, the temperature decreases. Depth and temperature are inversely related. Especially between depths of 2 to 6 meters, the distribution of observed temperatures ranges the widest - the distribution may be up to +/- 10 degrees from the linear model. Additionally, there are many more data points for depths under 7 meters, and much less data collected above 7 meters. The distribution of this data suggests that perhaps there is not a linear relationship between these variables.

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

```r
#7
# regression using lm function for temp by depth relationship
depth.temp.regression <-
  lm(data = NTL.LTER.processed, temperature_C ~ depth)

# summary of results of regression
summary(depth.temp.regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = NTL.LTER.processed)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -9.5077 -3.0182  0.0743  2.9248 13.6033
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.94872    0.06790   323.3   <2e-16 ***
## depth       -1.94700    0.01173  -166.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.829 on 9720 degrees of freedom
## Multiple R-squared:  0.7391, Adjusted R-squared:  0.7391
## F-statistic: 2.754e+04 on 1 and 9720 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: The coefficient for depth is -1.94, which tells us that when depth increases, the temperature decreases. The r-squared value is 0.7391, which means that 73.9% of the values of temperature are explained by depth. The degrees of freedom for this finding is 9720. For every one 1m increase in depth, the temperature will drop by -1.94 degrees C. The p value is less than the confidence level of 0.05, so we can determine that the coefficient is significant.

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
# creating a linear model with all three potential explanatory variables, start of AIC
JulyAIC <- lm(data = NTL.LTER.processed, temperature_C ~ year4 + daynum + depth)

# running stepwise algorithim to determine set of explanatory variables
step(JulyAIC)
```

```
## Start:  AIC=26016.31
## temperature_C ~ year4 + daynum + depth
##
##          Df Sum of Sq    RSS   AIC
## <none>                 141118 26016
## - year4   1        80 141198 26020
## - daynum  1      1333 142450 26106
## - depth   1    403925 545042 39151
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL.LTER.processed)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
```

```
##    -6.45556      0.01013      0.04134     -1.94726
```

```
#10

# running multiple linear regression with year4, daynum, and depth as variables & summary
Temp.model <-
  lm(data = NTL.LTER.processed, temperature_C ~
       year4 + daynum + depth)
summary(Temp.model)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL.LTER.processed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6517 -2.9937  0.0855  2.9692 13.6171
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -6.455560   8.638808   -0.747   0.4549
## year4        0.010131   0.004303    2.354   0.0186 *
## daynum       0.041336   0.004315    9.580   <2e-16 ***
## depth       -1.947264   0.011676 -166.782   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.811 on 9718 degrees of freedom
## Multiple R-squared:  0.7417, Adjusted R-squared:  0.7417
## F-statistic:  9303 on 3 and 9718 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 AIC method suggests using year4, daynum, and depth to predict the temperature in the multiple regression. Removing any of these three variables would increase the AIC. The R squared value is 0.7417 so 74% of the observed variance of temperature can be explained by this model. Technically it is an increase in 0.26% higher than the model using only depth for the explanation of the observed variance. However, I would argue that this model is over-parameterized and not an improvement because adding more variables is making the model more complex, changing the degrees of freedom, and the 0.26% is not worth it.

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
# ANOVA models with aov function
Temp.lakename.anova <-
  aov(data = NTL.LTER.processed, temperature_C ~ lakename)
summary(Temp.lakename.anova)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## lakename       8  21214  2651.8   49.04 <2e-16 ***
## Residuals   9713 525188    54.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#linear models running ANOVA test
Temp.lake.anova2 <-
  lm(data = NTL.LTER.processed, temperature_C ~ lakename)
summary(Temp.lake.anova2)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = NTL.LTER.processed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.766  -6.592  -2.692   7.634  23.832
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 17.6731     0.6741  26.218  < 2e-16 ***
## lakenameCrampton Lake       -2.3212     0.7902  -2.938  0.00332 **
## lakenameEast Long Lake      -7.4054     0.7143 -10.367  < 2e-16 ***
## lakenameHummingbird Lake    -6.8998     0.9594  -7.192 6.88e-13 ***
## lakenamePaul Lake           -3.8813     0.6891  -5.633 1.82e-08 ***
## lakenamePeter Lake          -4.3710     0.6878  -6.355 2.18e-10 ***
## lakenameTuesday Lake        -6.6073     0.7002  -9.437  < 2e-16 ***
## lakenameWard Lake           -3.2145     0.9594  -3.350  0.00081 ***
## lakenameWest Long Lake      -6.0876     0.7115  -8.556  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.353 on 9713 degrees of freedom
## Multiple R-squared:  0.03883,    Adjusted R-squared:  0.03803
## F-statistic: 49.04 on 8 and 9713 DF,  p-value: < 2.2e-16
```

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

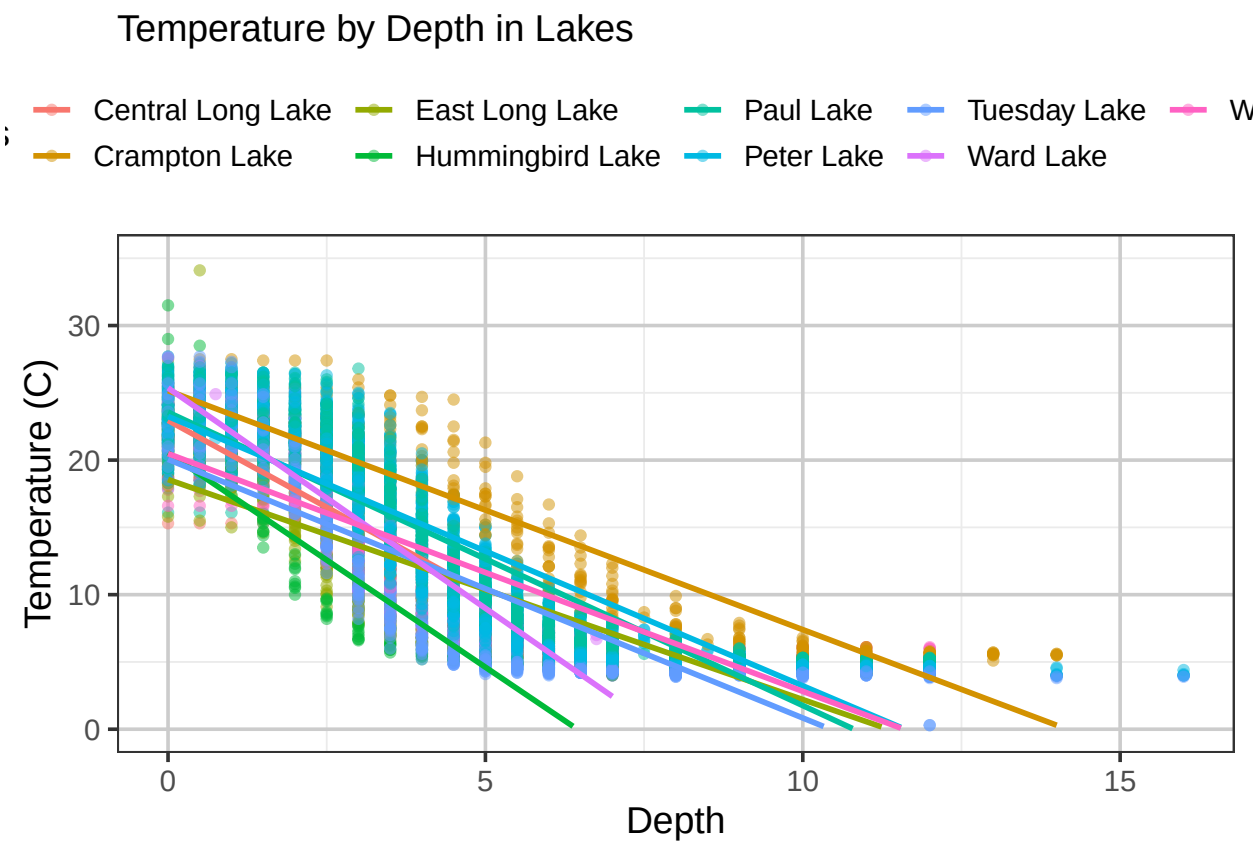
    Answer: Yes, there is significant difference in the mean temperature among the lakes. The estimates for each lake range between -7.4 and 17.67 (the intercept), and the p value for each lake is less than 0.05 so we can reject the null hypothesis. As a result, we reject the null hypothesis that all the mean temperatures among the lakes are the same.

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.
# Creating a scatterplot to look at depth and temp by lake
Plot.by.lake <-
  ggplot(NTL.LTER.processed,
         aes(x = depth, y = temperature_C, color = lakename)) +
geom_point(alpha = 0.5) +
geom_smooth(method = "lm", se = FALSE) +
  ylim(0, 35)+
  labs(x ="Depth", y ="Temperature (C)", color = "Lakes") +
```

```
  ggtitle("Temperature by Depth in Lakes")
print(Plot.by.lake)
```

```
## `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
# Running Tukey HSD for object derived from function aov
TukeyHSD(Temp.lakename.anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = NTL.LTER.processed)
##
## $lakename
##                                         diff        lwr        upr     p adj
## Crampton Lake-Central Long Lake    -2.3212225 -4.7727515  0.1303066 0.0801309
## East Long Lake-Central Long Lake   -7.4054440 -9.6215318 -5.1893561 0.0000000
## Hummingbird Lake-Central Long Lake -6.8998334 -9.8763946 -3.9232722 0.0000000
## Paul Lake-Central Long Lake        -3.8813120 -6.0191419 -1.7434822 0.0000007
## Peter Lake-Central Long Lake       -4.3710346 -6.5048955 -2.2371736 0.0000000
## Tuesday Lake-Central Long Lake     -6.6072831 -8.7795517 -4.4350145 0.0000000
## Ward Lake-Central Long Lake        -3.2144886 -6.1910498 -0.2379273 0.0229685
## West Long Lake-Central Long Lake   -6.0875867 -8.2949346 -3.8802388 0.0000000
```

```
## East Long Lake-Crampton Lake       -5.0842215 -6.5587481 -3.6096949 0.0000000
## Hummingbird Lake-Crampton Lake     -4.5786109 -7.0531008 -2.1041211 0.0000003
## Paul Lake-Crampton Lake            -1.5600896 -2.9141574 -0.2060217 0.0106305
## Peter Lake-Crampton Lake           -2.0498121 -3.3976050 -0.7020192 0.0000841
## Tuesday Lake-Crampton Lake         -4.2860606 -5.6938725 -2.8782488 0.0000000
## Ward Lake-Crampton Lake            -0.8932661 -3.3677559  1.5812237 0.9713958
## West Long Lake-Crampton Lake       -3.7663643 -5.2277226 -2.3050060 0.0000000
## Hummingbird Lake-East Long Lake     0.5056106 -1.7358512  2.7470723 0.9988025
## Paul Lake-East Long Lake            3.5241319  2.6670727  4.3811912 0.0000000
## Peter Lake-East Long Lake           3.0344094  2.1872987  3.8815201 0.0000000
## Tuesday Lake-East Long Lake         0.7981609 -0.1415120  1.7378337 0.1721160
## Ward Lake-East Long Lake            4.1909554  1.9494937  6.4324171 0.0000002
## West Long Lake-East Long Lake       1.3178572  0.2997124  2.3360021 0.0019544
## Paul Lake-Hummingbird Lake          3.0185213  0.8543999  5.1826428 0.0005172
## Peter Lake-Hummingbird Lake         2.5287988  0.3685979  4.6889997 0.0086420
## Tuesday Lake-Hummingbird Lake       0.2925503 -1.9055981  2.4906986 0.9999773
## Ward Lake-Hummingbird Lake          3.6853448  0.6898445  6.6808451 0.0043115
## West Long Lake-Hummingbird Lake     0.8122467 -1.4205745  3.0450678 0.9700210
## Peter Lake-Paul Lake               -0.4897225 -1.1036180  0.1241730 0.2442990
## Tuesday Lake-Paul Lake             -2.7259711 -3.4623514 -1.9895907 0.0000000
## Ward Lake-Paul Lake                 0.6668235 -1.4972980  2.8309450 0.9895659
## West Long Lake-Paul Lake           -2.2062747 -3.0404749 -1.3720745 0.0000000
## Tuesday Lake-Peter Lake            -2.2362485 -2.9610258 -1.5114713 0.0000000
## Ward Lake-Peter Lake                1.1565460 -1.0036549  3.3167469 0.7703831
## West Long Lake-Peter Lake          -1.7165522 -2.5405279 -0.8925764 0.0000000
## Ward Lake-Tuesday Lake              3.3927945  1.1946462  5.5909429 0.0000597
## West Long Lake-Tuesday Lake         0.5196964 -0.3991749  1.4385677 0.7121762
## West Long Lake-Ward Lake           -2.8730982 -5.1059193 -0.6402770 0.0021521
```

```
# Groups for pairwise
Lakes.HSD.groups <-
  HSD.test(Temp.lakename.anova, "lakename", group = TRUE)
Lakes.HSD.groups
```

```
## $statistics
##    MSerror   Df     Mean       CV
##   54.07064 9713 12.70646 57.87035
##
## $parameters
##    test   name.t ntr StudentizedRange alpha
##   Tukey lakename   9         4.387505  0.05
##
## $means
##                    temperature_C      std    r        se Min  Max    Q25   Q50
## Central Long Lake      17.67311 4.273404  119 0.6740735 8.9 26.8 14.400 18.40
## Crampton Lake          15.35189 7.244773  318 0.4123511 5.0 27.5  7.525 16.90
## East Long Lake         10.26767 6.766804  968 0.2363432 4.2 34.1  4.975  6.50
## Hummingbird Lake       10.77328 7.017845  116 0.6827343 4.0 31.5  5.200  7.00
## Paul Lake              13.79180 7.291951 2643 0.1430317 4.7 27.7  6.500 12.40
## Peter Lake             13.30207 7.667550 2892 0.1367356 4.0 27.0  5.600 11.40
## Tuesday Lake           11.06583 7.694274 1507 0.1894192 0.3 27.7  4.400  6.80
## Ward Lake              14.45862 7.409079  116 0.6827343 5.7 27.6  7.200 12.55
## West Long Lake         11.58552 6.963995 1043 0.2276872 4.0 25.7  5.400  8.00
##                       Q75
## Central Long Lake 21.350
```

```
## Crampton Lake     22.300
## East Long Lake    15.925
## Hummingbird Lake  15.625
## Paul Lake         21.400
## Peter Lake        21.500
## Tuesday Lake      19.400
## Ward Lake         23.200
## West Long Lake    18.800
##
## $comparison
## NULL
##
## $groups
##                   temperature_C groups
## Central Long Lake      17.67311      a
## Crampton Lake          15.35189     ab
## Ward Lake              14.45862     bc
## Paul Lake              13.79180      c
## Peter Lake             13.30207      c
## West Long Lake         11.58552      d
## Tuesday Lake           11.06583     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, Paul Lake and Ward Lake have the same mean temperature. They are all in group "c". There is not a group with only one lake in it, so there is no lake with a mean temperature that is 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: If we are looking at just Peter and Paul Lakes, we might try a two sample T-test to test the hypothesis that the means are equivalent. The alternate hypothesis would be that the the mean temperatures of Peter and Paul Lakes are statistically distinct.

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?

```
# processing data to be only Crampton and Ward Lakes
Crampton.Ward.July <-
  NTL.LTER.processed %>%
  filter(lakename == "Crampton Lake" | lakename == "Ward Lake")

# running a two sample t test:
Crampton.Ward.twosamplet <-
  t.test(Crampton.Ward.July$temperature_C ~ Crampton.Ward.July$lakename)
Crampton.Ward.twosamplet
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  Crampton.Ward.July$temperature_C by Crampton.Ward.July$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
```

Answer: The p value of the two sample t test is 0.265, which is higher than the 0.05 confidence level, so we cannot reject the null hypothesis. Crampton and Ward Lakes, according to this test, may have the same mean temperatures. According the results found in the HSD groups in Questions 15 and 16, both Crampton and Ward Lakes were in group b, so they have mean temperatures that are not statistically distinct.