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

David Liddle

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 Set up session
#load packages
library(tidyverse);library(lubridate);library(here);library(ggplot2);
library(cowplot);library(knitr); library(agricolae);library(dplyr)
#Get working directory
getwd()
```

```
## [1] "/Users/davidliddle/Documents/EDA-Spring2023"
```

```
#Import raw data
lakechem <- read.csv(
  here("data", "raw", "NTL-LTER_Lake_ChemistryPhysics_Raw.csv"),
  stringsAsFactors = TRUE) %>%
mutate(sampledate = mdy(sampledate))
#2 Build ggplot theme
david_theme <- theme_classic(base_size = 14) +
theme(axis.text = element_text(color = "black"),
legend.position = "right",
```

```
panel.grid.minor = element_line(color = "gray", linetype = "solid"),
panel.grid.major = element_line(color = "gray", linetype = "solid"),
legend.background = element_rect(fill = "gray"))
#Set the theme
theme_set(david_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:The mean lake temperature recorded in July does not change with depth across all lakes. Ha:The mean lake temperature recorded in July changes 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)

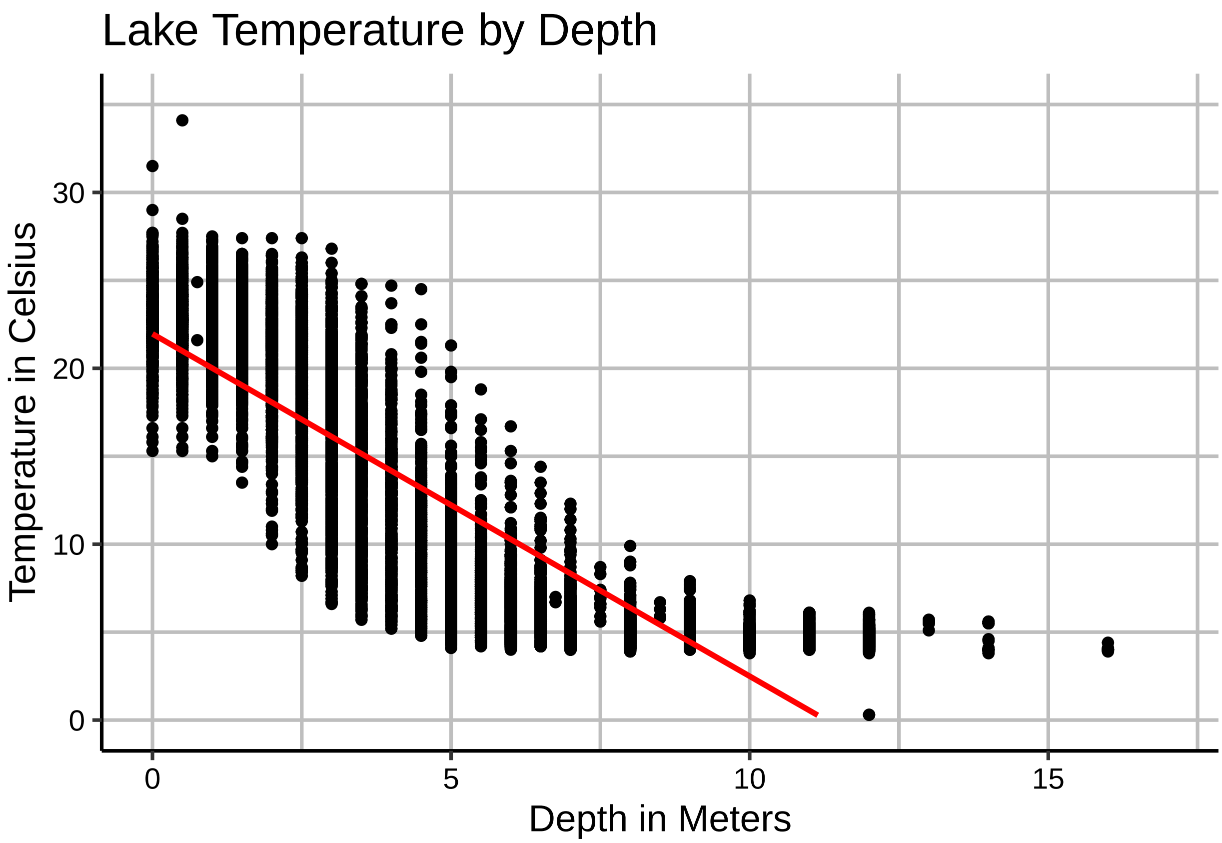
5. Visualize the relationship among the two continuous variables with a scatter plot of temperature by depth. Add a smoothed line showing the linear model, and limit temperature values from 0 to 35 °C. Make this plot look pretty and easy to read.

```
#4 Wrangle NTL-LTER dataset
lakechem_select <- lakechem %>%
  filter(month(sampledate) == 7) %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  drop_na()

#5 Create scatter plot of temperature by depth
plot <- ggplot(lakechem_select, aes(x = depth, y = temperature_C)) +
# Add points for each data point
geom_point() +
geom_smooth(method = "lm", se = FALSE, col = "red") +
  labs(title = "Lake Temperature by Depth",
       x = "Depth in Meters",
       y = "Temperature in Celsius") + # Name axes
   xlim(c(0, 17)) + ylim(c(0, 35)) #Add x and y limits
print(plot)
```

```
## `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: The trendline on the scatterplot indicates that temperature is inversely correlated with depth, meaning that as the depth of the lake increases, the temperature decreases. The distribution of points do not appear to follow a linear pattern, suggesting that a single linear regression for all the lakes combined may not be appropriate.

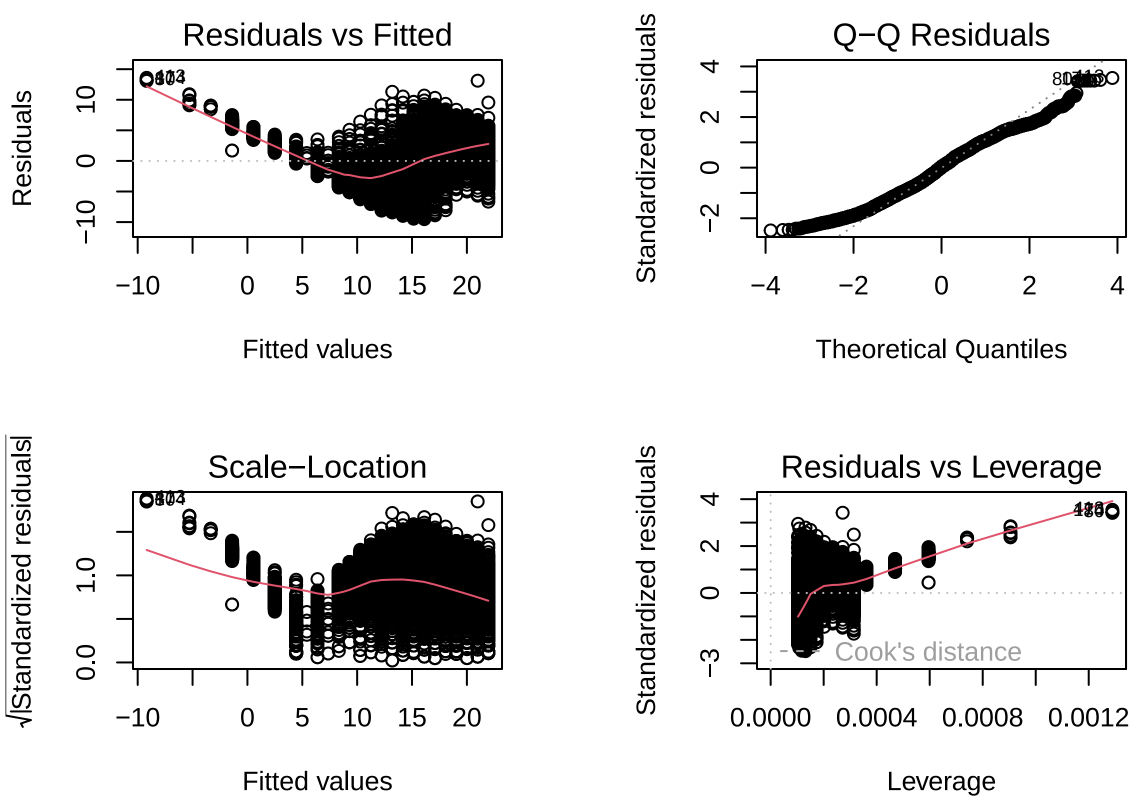
7. Perform a linear regression to test the relationship and display the results

```
#7 Perform linear regression test
tempbydepth <- lm(data = lakechem_select, temperature_C ~ depth)
#summary of Results
summary(tempbydepth)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = lakechem_select)
##
## 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
```

```r
#Plot linear regression
par(mfrow = c(2,2), mar=c(4,4,4,4))
plot(tempbydepth)
```




```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: The model calculated a p-value of < 0.05 (<2e-16), meaning that we can reject the null hypothesis that lake depth has no effect on the lake temperature. A Multiple R-squared value of 0.7387 indicates a strong correlataion between lake depth and temperature. The R-Squared value means that the change in depth explains approximately 73.87% of the variability in lake temperature. For this model we had 9,726 degrees of freedom. For everyone one meter change in depth, temperature is expected to decrease by 1.95 degrees 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 Create AIC to determine what set of variables is the best temperature predictor
TPAIC <- lm(data = lakechem_select, temperature_C ~ depth + daynum + year4)
step(TPAIC)
```

```
## 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 = lakechem_select)
##
## Coefficients:
## (Intercept)        depth       daynum        year4
##    -8.57556     -1.94644      0.03978      0.01134
```

```
#10 Define the multiple regression model
model <- lm(temperature_C ~ year4 + daynum + depth, data = lakechem_select)

# Fit the model
summary(model)
```

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

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 that we use all of the variables (year4, daynum, and depth) to predict temperature in our multiple regression. This is because when none of the above variables are removed, the AIC value is the lowest. This is an improvement over using only depth as the explanatory variable. When only using depth, the Multiple R-squared was 0.7387. When the other variables were factored in the Multiple R-squared rose to 0.7412. This value indicates a higher proportion of the observed variance is explained by the models with the addtional variables.

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).

```r
#12 Run ANOVA test

# Format ANOVA as aov
lakechem_select.anova <- aov(data = lakechem_select, temperature_C ~ lakename)
summary(lakechem_select.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
```

```r
#results: reject null hypothesis i.e. difference
#between a pair of group means is statistically significant (for a < 0.05)

# Format ANOVA as lm
lakechem_select.anova2 <- lm(data = lakechem_select, temperature_C ~ lakename)
summary(lakechem_select.anova2)
```

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

```
#results: reject null hypothesis i.e. difference
#between a pair of group means is statistically significant (for a < 0.05)
```

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

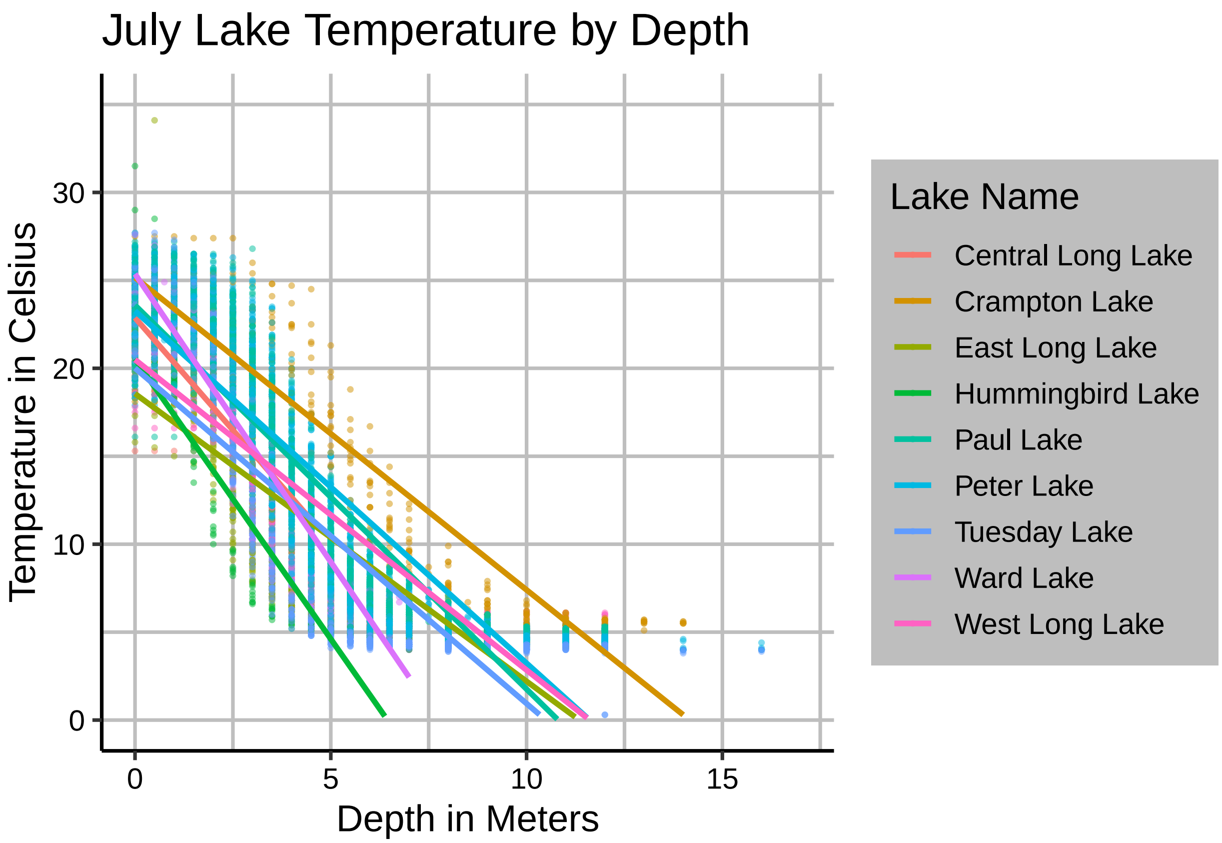
    Answer: The p-value associated with the F-statistic is <2e-16, which is extremely close to 0. Therefore, there is extremely strong evidence supporting that there is a difference in mean temperature among the lakes.

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. Create scatter plot of temperature by depth
plot2 <- ggplot(lakechem_select, aes(x = depth, y = temperature_C,
                                     color = lakename)) +
# Add points for each data point
geom_point(size = 0.5, alpha = 0.5) +
geom_smooth(method = "lm", se = FALSE) +
  labs(title = "July Lake Temperature by Depth",
       x = "Depth in Meters",
       y = "Temperature in Celsius",
       color = "Lake Name") + # Name axes
   xlim(c(0, 17)) + ylim(c(0, 35)) #Add x and y limits
print(plot2)
```

```
## `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 Post-hoc test
lakechem.tukey <- HSD.test(lakechem_select.anova, "lakename", group = TRUE)
lakechem.tukey
```

```
## $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        se Min  Max    Q25   Q50
## Central Long Lake       17.66641 4.196292  128 0.6501298 8.9 26.8 14.400 18.40
## Crampton Lake           15.35189 7.244773  318 0.4124692 5.0 27.5  7.525 16.90
## East Long Lake          10.26767 6.766804  968 0.2364108 4.2 34.1  4.975  6.50
## Hummingbird Lake        10.77328 7.017845  116 0.6829298 4.0 31.5  5.200  7.00
## Paul Lake               13.81426 7.296928 2660 0.1426147 4.7 27.7  6.500 12.40
## Peter Lake              13.31626 7.669758 2872 0.1372501 4.0 27.0  5.600 11.40
## Tuesday Lake            11.06923 7.698687 1524 0.1884137 0.3 27.7  4.400  6.80
## Ward Lake               14.45862 7.409079  116 0.6829298 5.7 27.6  7.200 12.55
## West Long Lake          11.57865 6.980789 1026 0.2296314 4.0 25.7  5.400  8.00
##                      Q75
```

```
## Central Long Lake 21.000
## 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      17.66641      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: Based on the results of the Tukey Test, the lakes with the same mean temperature as Peter Lake are Ward Lake and Paul Lake. There is no lake with a statistically distinct mean temperature from the other lakes. This is because every group letter is shared with at least one other lake. If, for example, there was a distinct lake, it would be grouped as 'f' and would not share this letter with any other lake.

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 use a two-tailed t-test to determine if they have statistically significant mean temperatures. A two-tailed t-test is used when you want to test for the significance of a difference without assuming a specific direction for that difference.

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
#Filter data for records only for Crampton and Ward Lake
lakechem_filtered <- lakechem_select %>%
```

```
  filter(lakename == c("Crampton Lake", "Ward Lake"))
#Run a two-sample T-test
#Format as a t-test
lakechem.twosample <- t.test(data = lakechem_filtered, temperature_C ~ lakename)
lakechem.twosample
```

```
##
##  Welch Two Sample t-test
##
## data:  temperature_C by lakename
## t = 0.98673, df = 95.77, p-value = 0.3263
## alternative hypothesis: true difference in means between group Crampton Lake and group Ward Lake is n
## 95 percent confidence interval:
##  -1.130614  3.365610
## sample estimates:
## mean in group Crampton Lake     mean in group Ward Lake
##                    15.37107                    14.25357
```

> Answer: The results of the two-tailed test resulted in a p-value of p = 0.3263, which is not statistically significant for an alpha = 0.05. Therefore, we do not reject the null hypothesis that the mean temperatures between Crampton and Ward Lake are equal. The mean temperatures for each lake are not equal, but are too close together to infer significance. This does match our answer from part 16, because Crampton Lake is part of group 'ab' and Ward Lake is part of group 'bc'. This means that they share commonailty in group 'b', meaning that they do not fall into statistically separate classes.