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

Jordan Mullens

Spring 2023

## OVERVIEW

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

### Directions

1. Rename this file `JordanMullens_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
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.0
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.2.1     v stringr 1.5.0
## v readr   2.1.3     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(agricolae)
library(here)
```

```
## here() starts at /home/guest/EDA-Spring2023
```

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

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

```
here()
```

```
## [1] "/home/guest/EDA-Spring2023"
```

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

NTLLTR.raw$sampledate <- myd(NTLLTR.raw$sampledate)
```

```
## Warning: 23670 failed to parse.
```

```
class(NTLLTR.raw$sampledate)
```

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

```
#2
library(ggthemes)
my_theme <- theme_base() +
  theme(
    line = element_line(
      color='black',
      linewidth =2
    ),
    legend.background = element_rect(
      color='grey',
      fill = 'green'
    ),
    legend.title = element_text(
      color='blue'
    ))
theme_set(my_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: Mean lake temperature is the same across lakes and depths Ha: Mean lake temperature is different across lakes and depths

4. Wrangle your NTL-LTER dataset with a pipe function so that the records meet the following criteria:

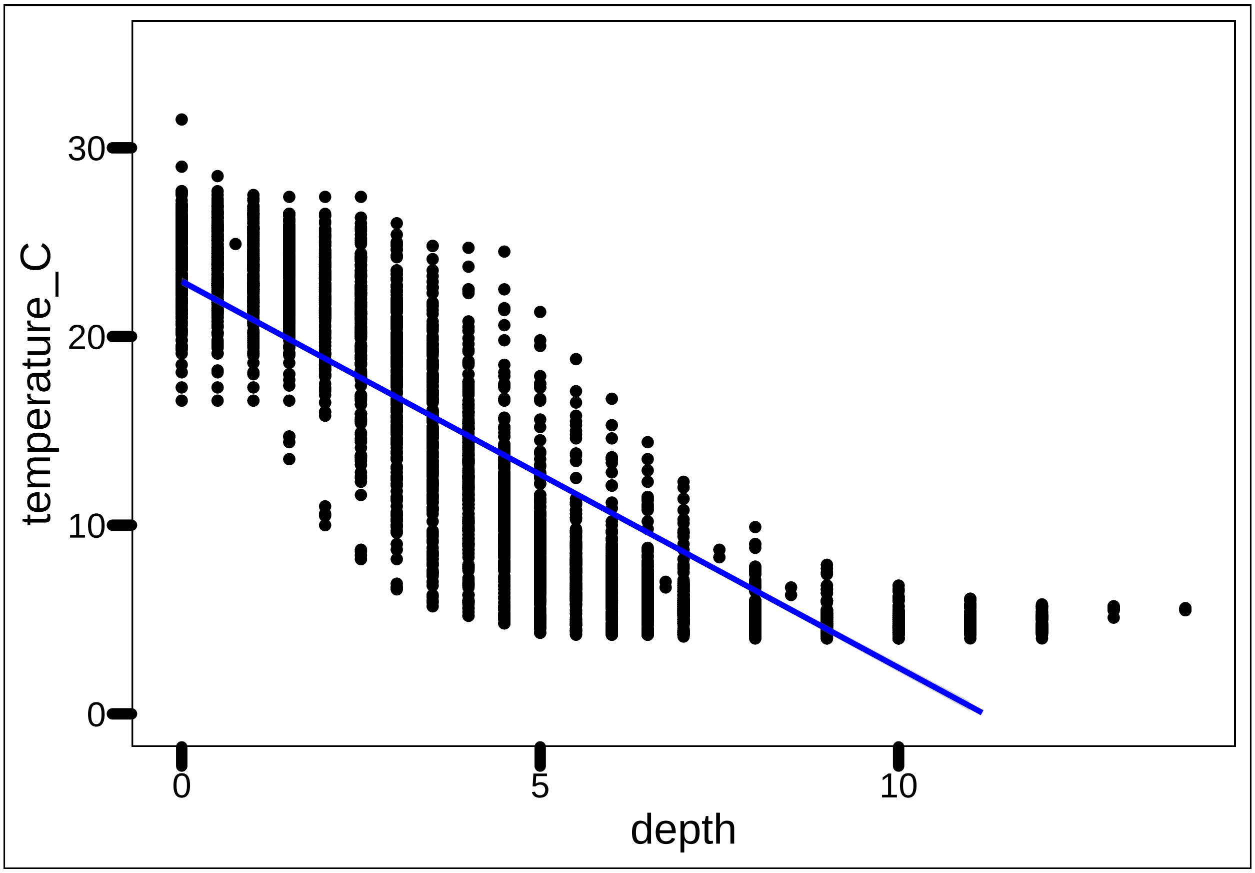
- Only dates in July.
- Only the columns: `lakename`, `year4`, `daynum`, `depth`, `temperature_C`
- Only complete cases (i.e., remove NAs)

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
NTLLTR.raw <- mutate(NTLLTR.raw, month = month(sampledate))


NTLLTR.wrangled <- NTLLTR.raw %>%
  filter(month == 7) %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  drop_na(temperature_C)


#5
NTLLTR.Temp.plot1 <-
  ggplot(NTLLTR.wrangled, aes(x = depth, y = temperature_C)) +
   geom_point() +
  geom_smooth(method = "lm", color="blue") +
  ylim(0, 35) +
  my_theme
print(NTLLTR.Temp.plot1)
```

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

```
## Warning: Removed 16 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?

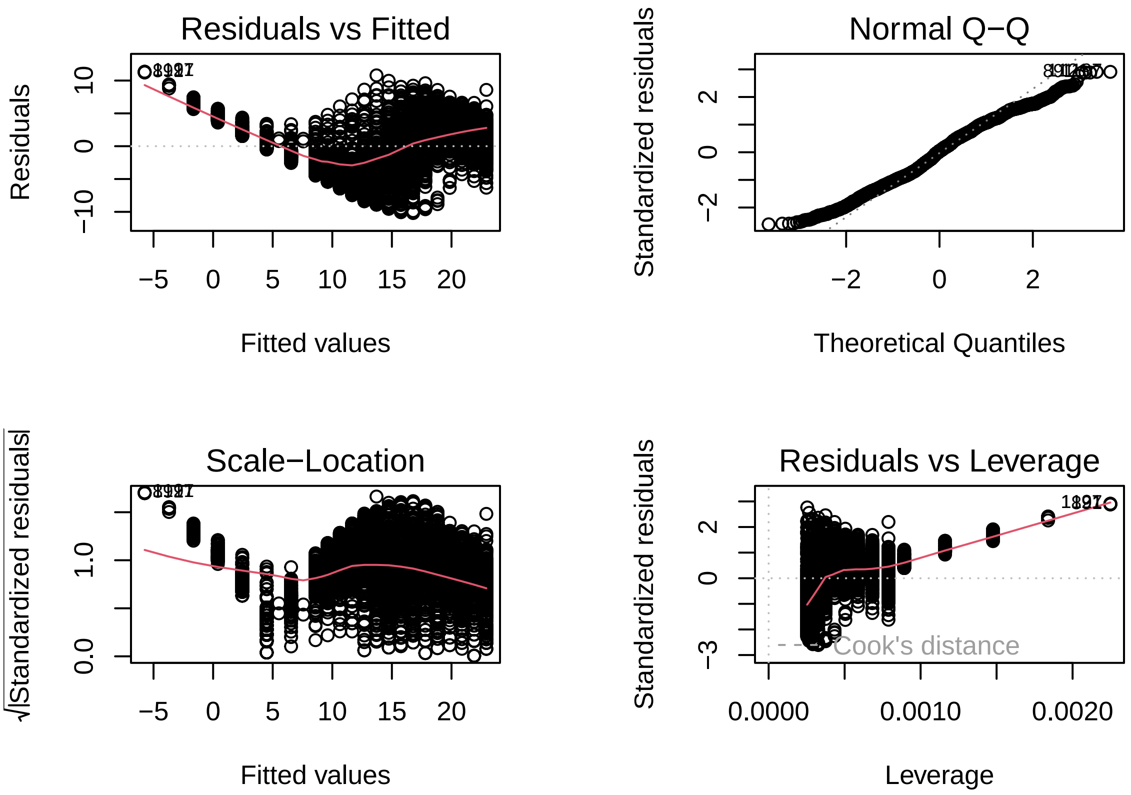
   Answer: It suggests temperature decreases as depth increases. Upon first inspection, the relationship could be linear from depths of 0 to about 8 meters. An exponential function may fit the data better.

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

```
#7
NTLLTR.Temp.Regression <- lm(data = NTLLTR.wrangled, temperature_C ~ depth)
summary(NTLLTR.Temp.Regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = NTLLTR.wrangled)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1805  -3.1610   0.1718   2.9479  11.3450
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.92375    0.10960   209.2   <2e-16 ***
## depth       -2.04777    0.01891  -108.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.903 on 3963 degrees of freedom
## Multiple R-squared:  0.7474, Adjusted R-squared:  0.7473
## F-statistic: 1.172e+04 on 1 and 3963 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2), mar=c(4,4,4,4))
plot(NTLLTR.Temp.Regression)
```



```
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 test has 9726 degrees of freedom. The results suggest that we reject the null hypothesis. Temperature is predicted to change 1.94 degrees C per 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
Temp.AIC <- lm(data = NTLLTR.wrangled, temperature_C ~ depth + year4 +
              daynum)
summary(Temp.AIC)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = NTLLTR.wrangled)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7422  -3.1863   0.1256   3.0337  11.5568
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 161.843864  26.080340    6.206 6.01e-10 ***
## depth        -2.048790   0.018758 -109.222  < 2e-16 ***
## year4        -0.073314   0.012965   -5.655 1.67e-08 ***
## daynum        0.042099   0.006783    6.207 5.96e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.869 on 3961 degrees of freedom
## Multiple R-squared:  0.7518, Adjusted R-squared:  0.7516
## F-statistic:  4000 on 3 and 3961 DF,  p-value: < 2.2e-16
```

```
#10
Temp.multi.regression <- lm(data = NTLLTR.wrangled, temperature_C ~ depth + daynum + year4)
summary(Temp.multi.regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + daynum + year4, data = NTLLTR.wrangled)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7422  -3.1863   0.1256   3.0337  11.5568
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 161.843864  26.080340    6.206 6.01e-10 ***
## depth        -2.048790   0.018758 -109.222  < 2e-16 ***
## daynum        0.042099   0.006783    6.207 5.96e-10 ***
## year4        -0.073314   0.012965   -5.655 1.67e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.869 on 3961 degrees of freedom
## Multiple R-squared:  0.7518, Adjusted R-squared:  0.7516
## F-statistic:  4000 on 3 and 3961 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: We could use depth, daynum, and year. However, daynum and depth appear to be stronger predictors of temperature than year.

---

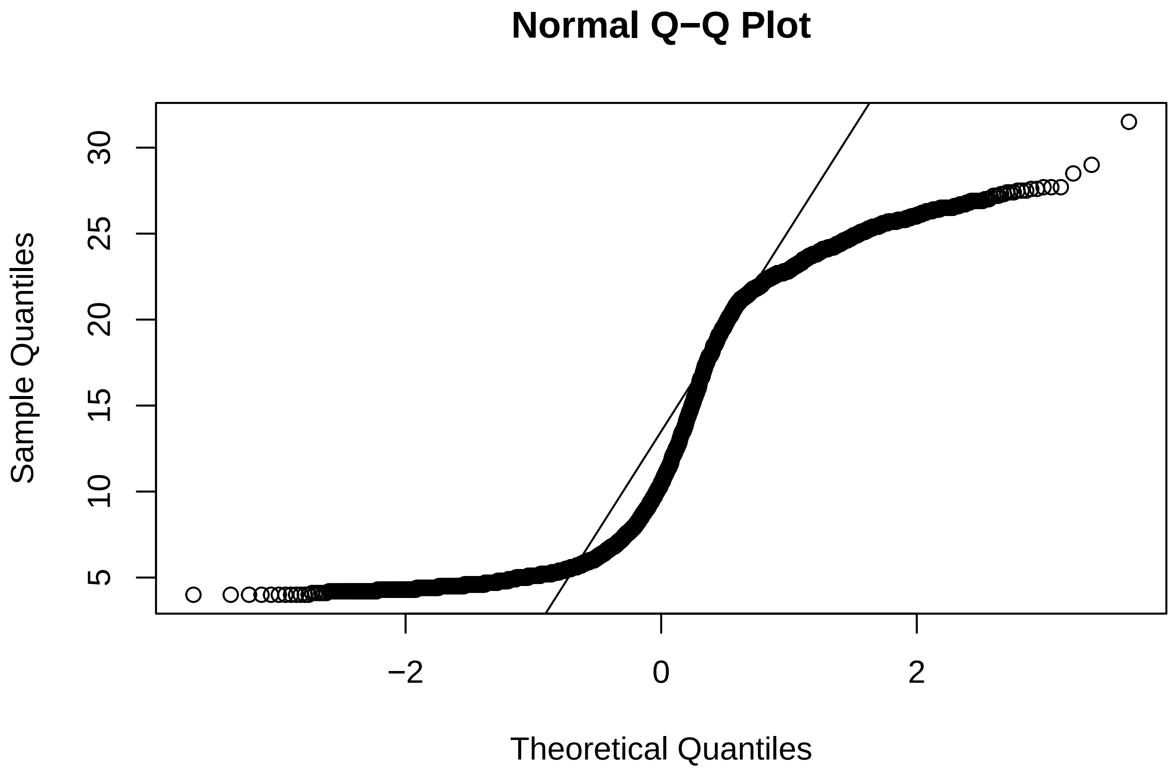
## Analysis of Variance

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
# Wrangle the data
Temp.Different.Lakes <- NTLLTR.wrangled %>%
  group_by(lakename, daynum, year4)

summary(Temp.Different.Lakes)
```

```
##             lakename        year4          daynum          depth
##  Peter Lake    :1400   Min.   :2001   Min.   :182.0   Min.   : 0.000
##  Paul Lake     :1370   1st Qu.:2004   1st Qu.:190.0   1st Qu.: 2.000
##  Tuesday Lake  : 500   Median :2008   Median :197.0   Median : 4.500
##  Crampton Lake : 298   Mean   :2008   Mean   :197.4   Mean   : 4.779
##  West Long Lake: 151   3rd Qu.:2013   3rd Qu.:205.0   3rd Qu.: 7.000
##  Ward Lake     : 116   Max.   :2016   Max.   :213.0   Max.   :14.000
##  (Other)       : 130
##  temperature_C
##  Min.   : 4.00
##  1st Qu.: 5.60
##  Median :10.50
##  Mean   :13.14
##  3rd Qu.:21.40
##  Max.   :31.50
##
```

```
#results: reject null in all except two: NIWO_057 and NIWO_046
#but method is robust from small deviations from normal distribution
qqnorm(NTLLTR.wrangled$temperature_C); qqline(NTLLTR.wrangled$temperature_C)
```



```
# Test for equal variance
# ANOVA is robust against departures from equal variance.
# bartlett.test() performs Bartlett's test of the null that the variances in each of the groups (sample
bartlett.test(Temp.Different.Lakes$temperature_C ~ Temp.Different.Lakes$lakename)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Temp.Different.Lakes$temperature_C by Temp.Different.Lakes$lakename
## Bartlett's K-squared = 10.597, df = 7, p-value = 0.1572
```

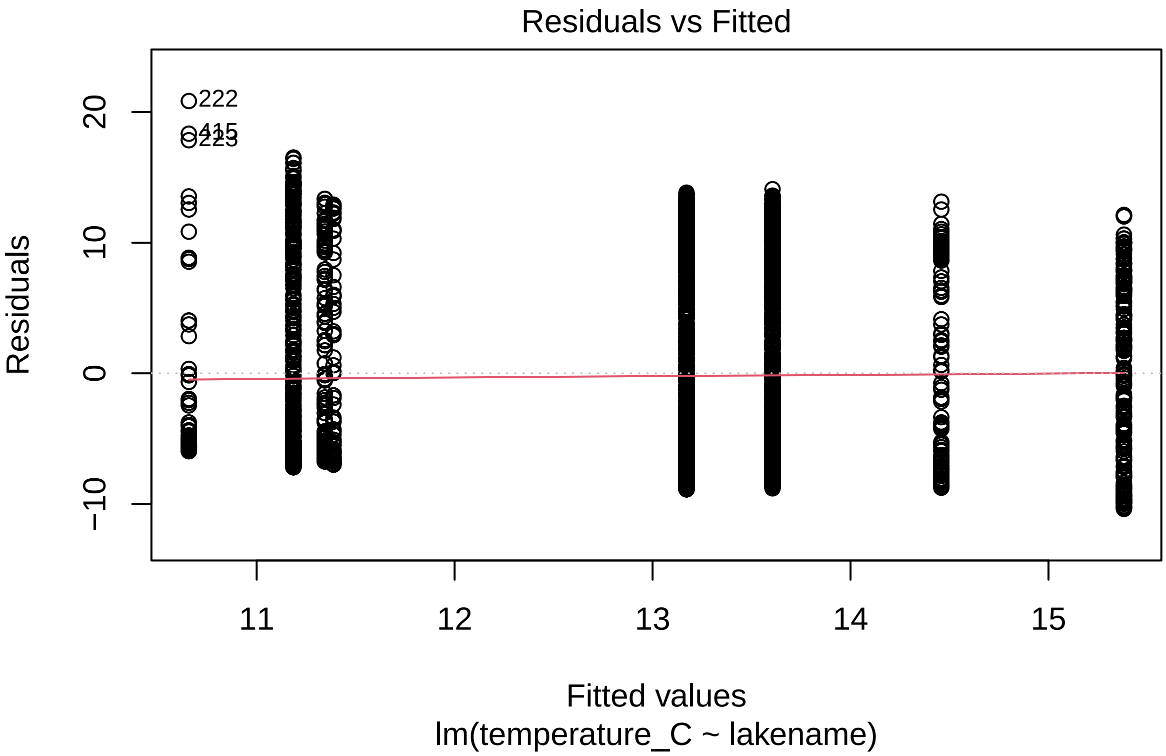
```
#results: reject null i.e. variances are not equal

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

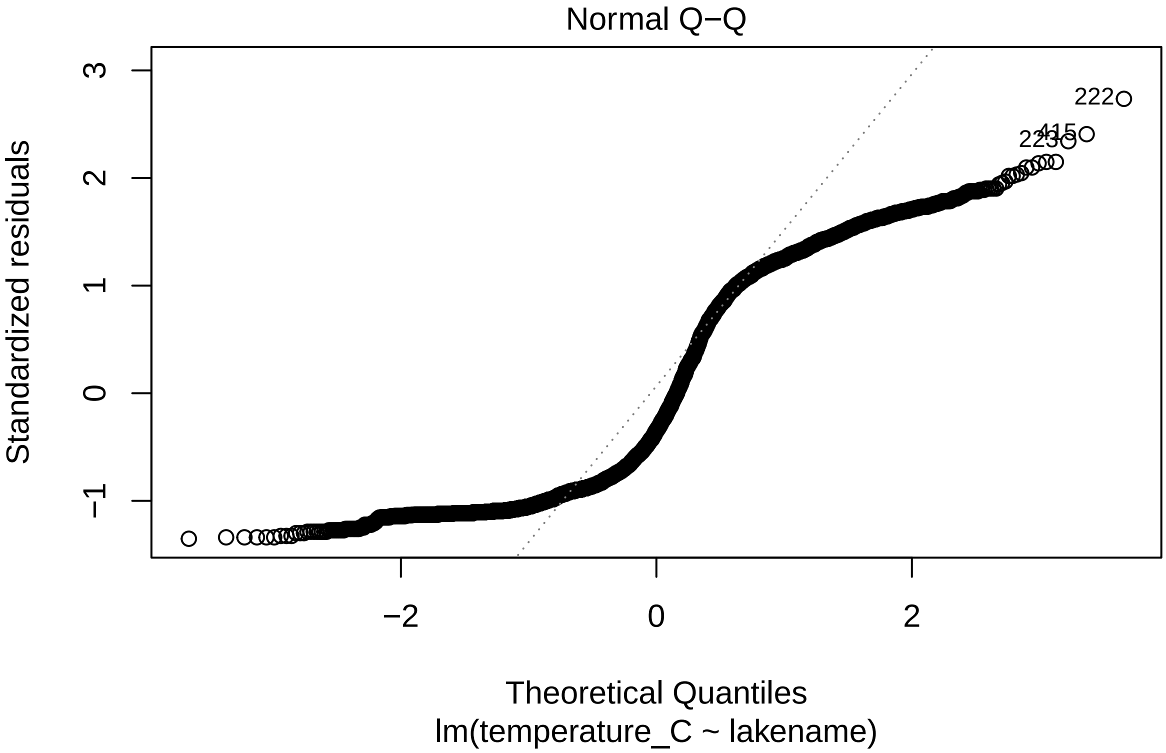
```
##                 Df Sum Sq Mean Sq F value  Pr(>F)
## lakename        7   4960   708.5   11.98 3.3e-15 ***
## Residuals    3957 234009    59.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

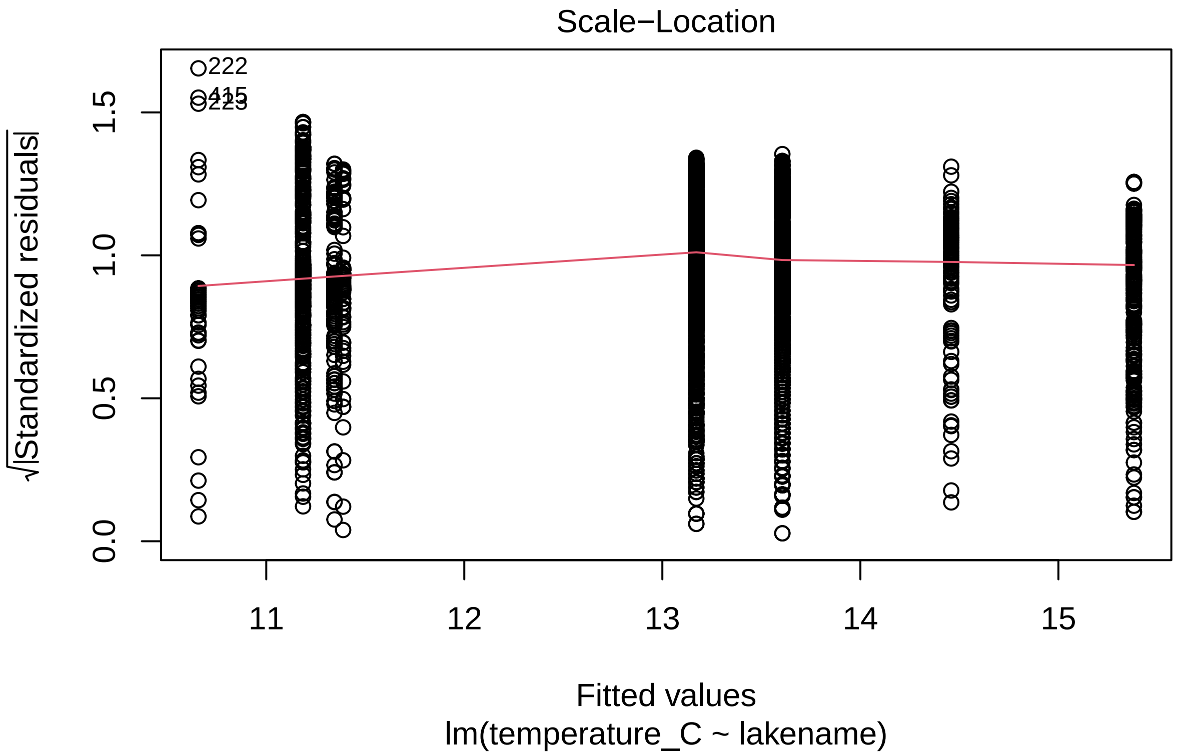
```
#results: reject null hypothesis i.e. difference between a pair of group means is statistically signifi

# Format ANOVA as lm
Temp.Different.Lakes.anova.2 <- lm(data = Temp.Different.Lakes, temperature_C ~ lakename)
summary(Temp.Different.Lakes.anova.2)
```

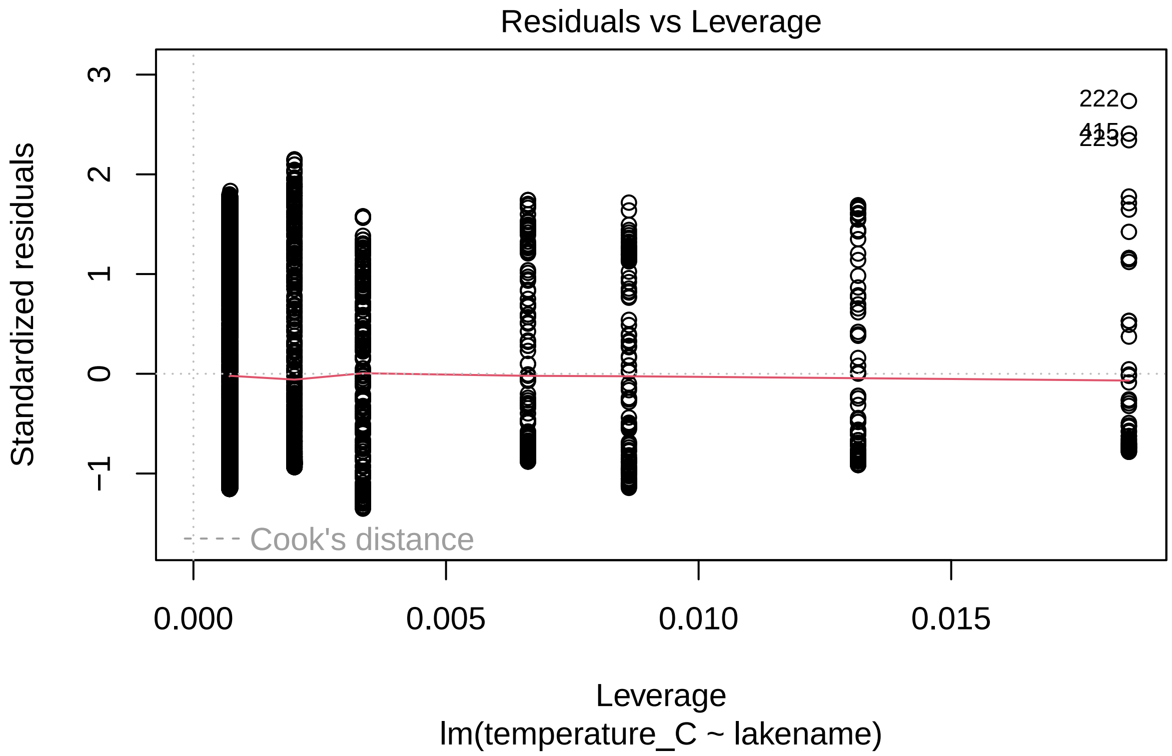
```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = Temp.Different.Lakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.381  -7.006  -2.681   8.028  20.843
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               15.3812     0.4455  34.528  < 2e-16 ***
## lakenameEast Long Lake    -3.9931     0.9882  -4.041 5.43e-05 ***
## lakenameHummingbird Lake  -4.7238     1.1374  -4.153 3.35e-05 ***
## lakenamePaul Lake         -1.7753     0.4915  -3.612 0.000308 ***
## lakenamePeter Lake        -2.2096     0.4906  -4.504 6.86e-06 ***
## lakenameTuesday Lake      -4.1954     0.5628  -7.455 1.10e-13 ***
## lakenameWard Lake         -0.9226     0.8416  -1.096 0.273034
## lakenameWest Long Lake    -4.0368     0.7682  -5.255 1.56e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.69 on 3957 degrees of freedom
## Multiple R-squared:  0.02075,    Adjusted R-squared:  0.01902
## F-statistic: 11.98 on 7 and 3957 DF,  p-value: 3.296e-15
```

```
# Checking model fit and assumptions
# ANOVA is robust against departures from normality.
plot(Temp.Different.Lakes.anova.2)
```


Residuals vs Fitted
Residuals
Fitted values
lm(temperature_C ~ lakename)
222
415
223
20
10
0
−10
11
12
13
14
15


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
0
2
222
415
223
Theoretical Quantiles
lm(temperature_C ~ lakename)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
222
415
223
11
12
13
14
15
Fitted values
lm(temperature_C ~ lakename)

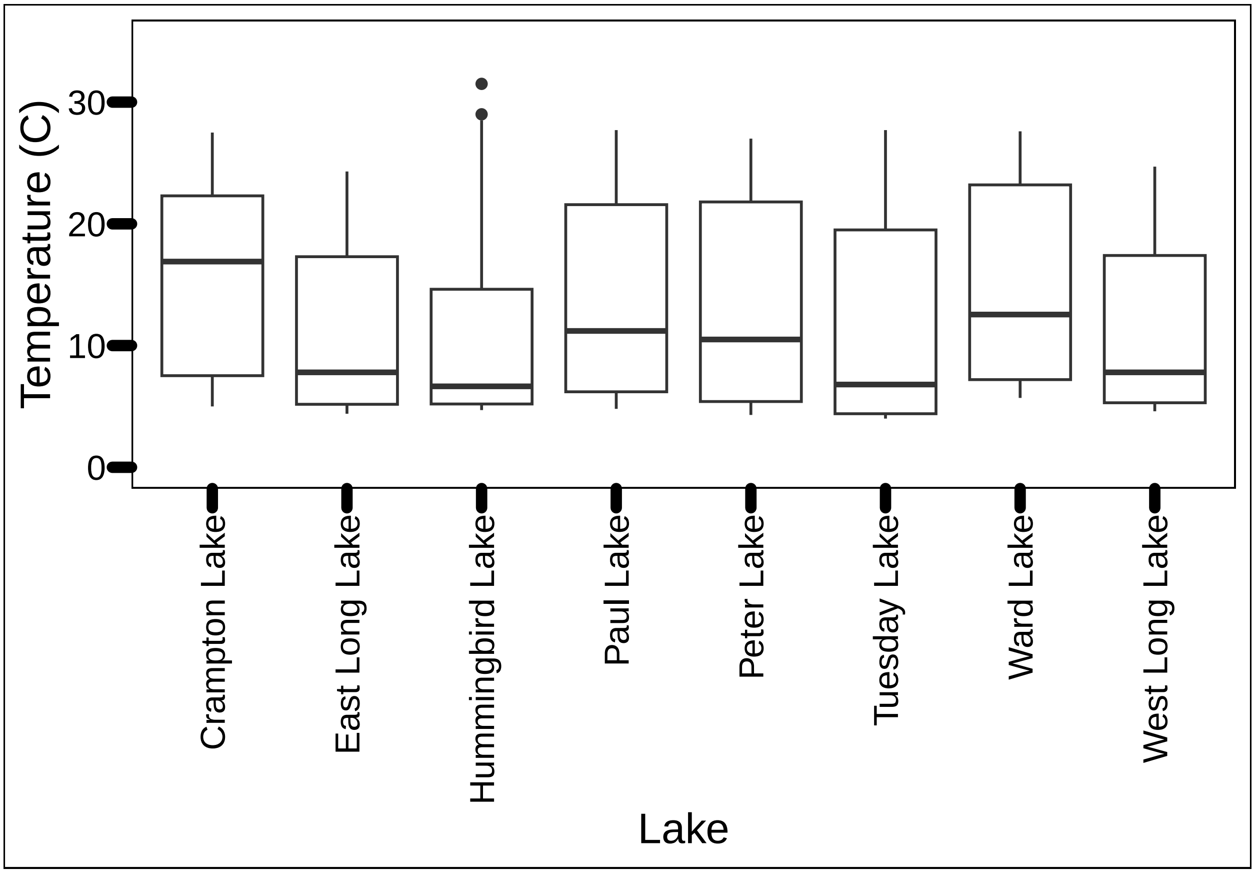


```
# Extract groupings for pairwise relationships
Temp.Different.Lakes.groups <- HSD.test(Temp.Different.Lakes.anova, "lakename", group = TRUE)
Temp.Different.Lakes.groups
```

```
## $statistics
##    MSerror   Df     Mean       CV
##   59.13787 3957 13.13695 58.53804
##
## $parameters
##    test   name.t ntr StudentizedRange alpha
##   Tukey lakename   8         4.288616  0.05
##
## $means
##                    temperature_C      std    r Min  Max   Q25   Q50    Q75
## Crampton Lake           15.38121 7.266740  298 5.0 27.5 7.525 16.90 22.300
## East Long Lake          11.38816 7.107479   76 4.4 24.3 5.175  7.80 17.300
## Hummingbird Lake        10.65741 7.526421   54 4.7 31.5 5.200  6.65 14.625
## Paul Lake               13.60593 7.572774 1370 4.8 27.7 6.200 11.20 21.575
## Peter Lake              13.17157 7.925438 1400 4.3 27.0 5.400 10.50 21.800
## Tuesday Lake            11.18580 7.963238  500 4.0 27.7 4.400  6.80 19.500
## Ward Lake               14.45862 7.409079  116 5.7 27.6 7.200 12.55 23.200
## West Long Lake          11.34437 6.926073  151 4.6 24.7 5.300  7.80 17.400
##
## $comparison
## NULL
##
```

```
## $groups
##                    temperature_C groups
## Crampton Lake           15.38121      a
## Ward Lake               14.45862     ab
## Paul Lake               13.60593      b
## Peter Lake              13.17157     bc
## East Long Lake          11.38816    bcd
## West Long Lake          11.34437     cd
## Tuesday Lake            11.18580      d
## Hummingbird Lake        10.65741      d
##
## attr(,"class")
## [1] "group"
```

```
# Graph the results
Temp.Different.Lakes.plot <- ggplot(Temp.Different.Lakes, aes(x = lakename, y = temperature_C)) +
  geom_boxplot() +
  labs(x = "Lake", y = "Temperature (C)") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5)) +
  ylim(0, 35)

print(Temp.Different.Lakes.plot)
```



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

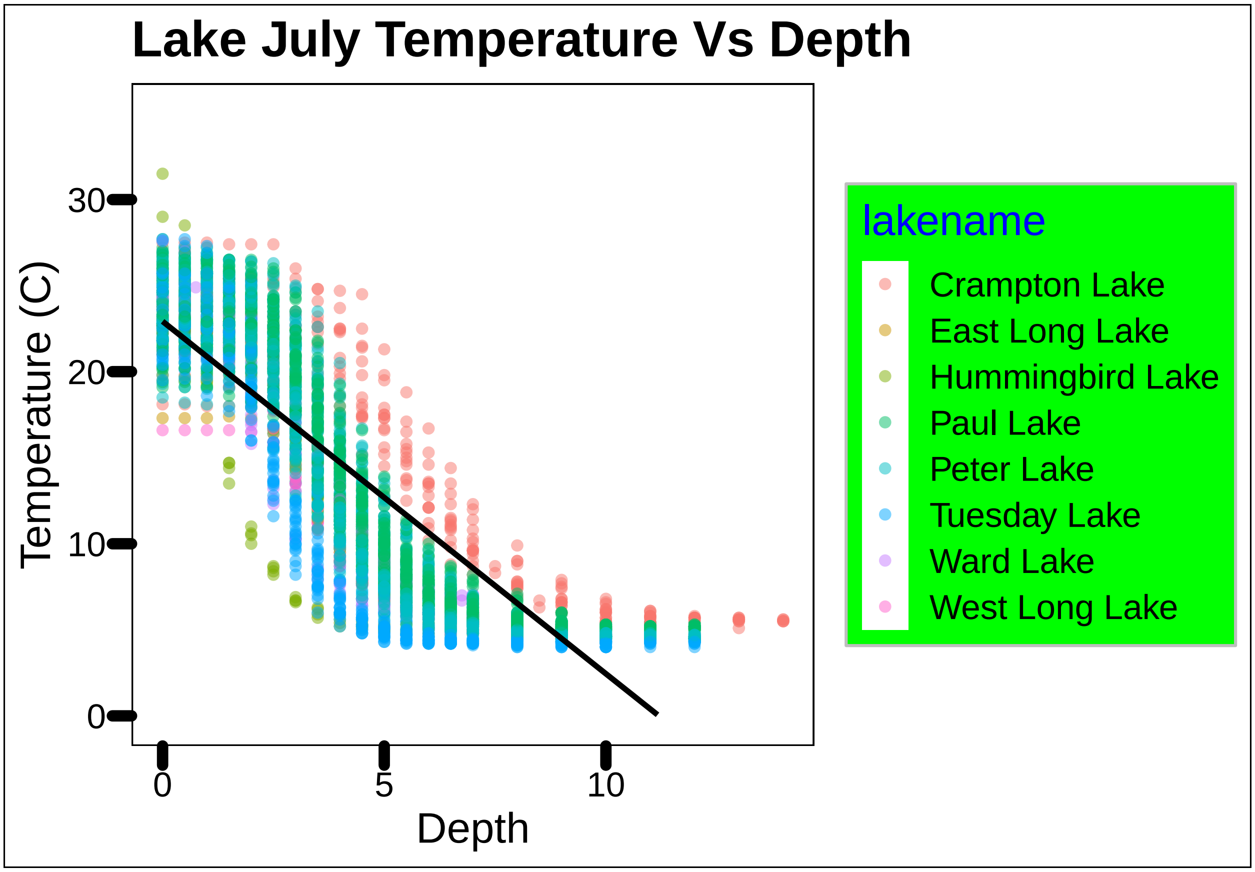
    Answer: Yes.

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.
Temp.Different.Lakes.Scatter <-
  ggplot(Temp.Different.Lakes, aes(
         x = depth,
         y = temperature_C,
         color = lakename)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method=lm, se=FALSE, color= "black") +
  ylim(0, 35) +
  labs(
    title = "Lake July Temperature Vs Depth",
    y= "Temperature (C)",
    x= "Depth",
    color= "lakename") +
  my_theme

print(Temp.Different.Lakes.Scatter)
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```

```
## Warning: Removed 16 rows containing missing values (‘geom_smooth()‘).
```



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

```
#15
# Post-hoc test
# TukeyHSD() computes Tukey Honest Significant Differences
TukeyHSD(Temp.Different.Lakes.anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = Temp.Different.Lakes)
##
## $lakename
##                                          diff        lwr        upr     p adj
## East Long Lake-Crampton Lake    -3.99305016 -6.9898372 -0.9962631 0.0014030
## Hummingbird Lake-Crampton Lake  -4.72380065 -8.1728653 -1.2747360 0.0008753
## Paul Lake-Crampton Lake         -1.77528105 -3.2658931 -0.2846690 0.0074364
## Peter Lake-Crampton Lake        -2.20963663 -3.6973926 -0.7218806 0.0001848
## Tuesday Lake-Crampton Lake      -4.19540805 -5.9020536 -2.4887625 0.0000000
## Ward Lake-Crampton Lake         -0.92258736 -3.4746895  1.6295148 0.9577614
## West Long Lake-Crampton Lake    -4.03683719 -6.3663311 -1.7073433 0.0000043
## Hummingbird Lake-East Long Lake -0.73075049 -4.8812753  3.4197743 0.9994810
## Paul Lake-East Long Lake         2.21776911 -0.5304546  4.9659928 0.2189853
## Peter Lake-East Long Lake        1.78341353 -0.9632621  4.5300892 0.5031662
## Tuesday Lake-East Long Lake     -0.20235789 -3.0734985  2.6687827 0.9999990
## Ward Lake-East Long Lake         3.07046279 -0.3710535  6.5119791 0.1210723
## West Long Lake-East Long Lake   -0.04378703 -3.3236271  3.2360531 1.0000000
## Paul Lake-Hummingbird Lake       2.94851960 -0.2869166  6.1839558 0.1045982
## Peter Lake-Hummingbird Lake      2.51416402 -0.7199574  5.7482854 0.2629225
## Tuesday Lake-Hummingbird Lake    0.52839259 -2.8120808  3.8688659 0.9997449
## Ward Lake-Hummingbird Lake       3.80121328 -0.0405765  7.6430031 0.0548734
## West Long Lake-Hummingbird Lake  0.68696345 -3.0106932  4.3846202 0.9992612
## Peter Lake-Paul Lake            -0.43435558 -1.3205940  0.4518829 0.8150592
## Tuesday Lake-Paul Lake          -2.42012701 -3.6385849 -1.2016691 0.0000001
## Ward Lake-Paul Lake              0.85269368 -1.4023509  3.1077383 0.9463593
## West Long Lake-Paul Lake        -2.26155615 -4.2611917 -0.2619205 0.0141456
## Tuesday Lake-Peter Lake         -1.98577143 -3.2007337 -0.7708092 0.0000206
## Ward Lake-Peter Lake             1.28704926 -0.9661085  3.5402070 0.6660344
## West Long Lake-Peter Lake       -1.82720057 -3.8247081  0.1703069 0.1018494
## Ward Lake-Tuesday Lake           3.27282069  0.8695020  5.6761394 0.0009675
## West Long Lake-Tuesday Lake      0.15857086 -2.0068984  2.3240402 0.9999987
## West Long Lake-Ward Lake        -3.11424983 -5.9934587 -0.2350409 0.0232984
```

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: Lakes with a p adj value above 0.05 have the same mean temperature as peter lake (Ward 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: We could use a two-tailed 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?

```
#Wrangling the data
NTLLTR.Crampton.Ward <- NTLLTR.wrangled %>%
  filter(lakename == "Crampton Lake" | lakename == "Ward Lake")

#Format as a t-test
#NTLLTR.Crampton.Ward$temperature_C will be our continuous dependent variable
#NTLLTR.Crampton.Ward$lakename will be our categorical variable with two levels (2018 and 2019)
NTLLTR.Crampton.Ward.twosample <- t.test(NTLLTR.Crampton.Ward$temperature_C ~ NTLLTR.Crampton.Ward$laken
NTLLTR.Crampton.Ward.twosample
```

```
##
##  Welch Two Sample t-test
##
## data:  NTLLTR.Crampton.Ward$temperature_C by NTLLTR.Crampton.Ward$lakename
## t = 1.144, df = 206.06, p-value = 0.254
## alternative hypothesis: true difference in means between group Crampton Lake and group Ward Lake is n
## 95 percent confidence interval:
##  -0.6674444  2.5126192
## sample estimates:
## mean in group Crampton Lake     mean in group Ward Lake
##                    15.38121                    14.45862
```

Answer: We accept the alternative hypothesis- there is a difference between Crampton Lake and Ward Lake.