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

Mandy Hooks

OVERVIEW

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

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Fay_A06_GLMs.Rmd") prior to submission.

The completed exercise is due on Tuesday, March 2 at 1:00 pm.

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] "/Users/mandyhooks/Environmental_Data_Analytics_2021/Assignments"
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(agricolae)
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

```
library(dplyr)

LakeChemPhys<-read.csv("/Users/mandyhooks/Environmental_Data_Analytics_2021/Data/Raw/NTL-LTER_Lake_Chem

LakeChemPhys<-
  LakeChemPhys %>%
  mutate(sampledate=mdy(sampledate))

LakeChemPhys$sampledate<-as.Date(LakeChemPhys$sampledate, format = "%m-%d-%Y")


#2
mytheme <- theme_classic(base_size = 20) +
  theme(axis.text = element_text(color = "black"),
        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: µ1 = µ2...µk Ha: µ1 != µ2...µk (not equal)

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)

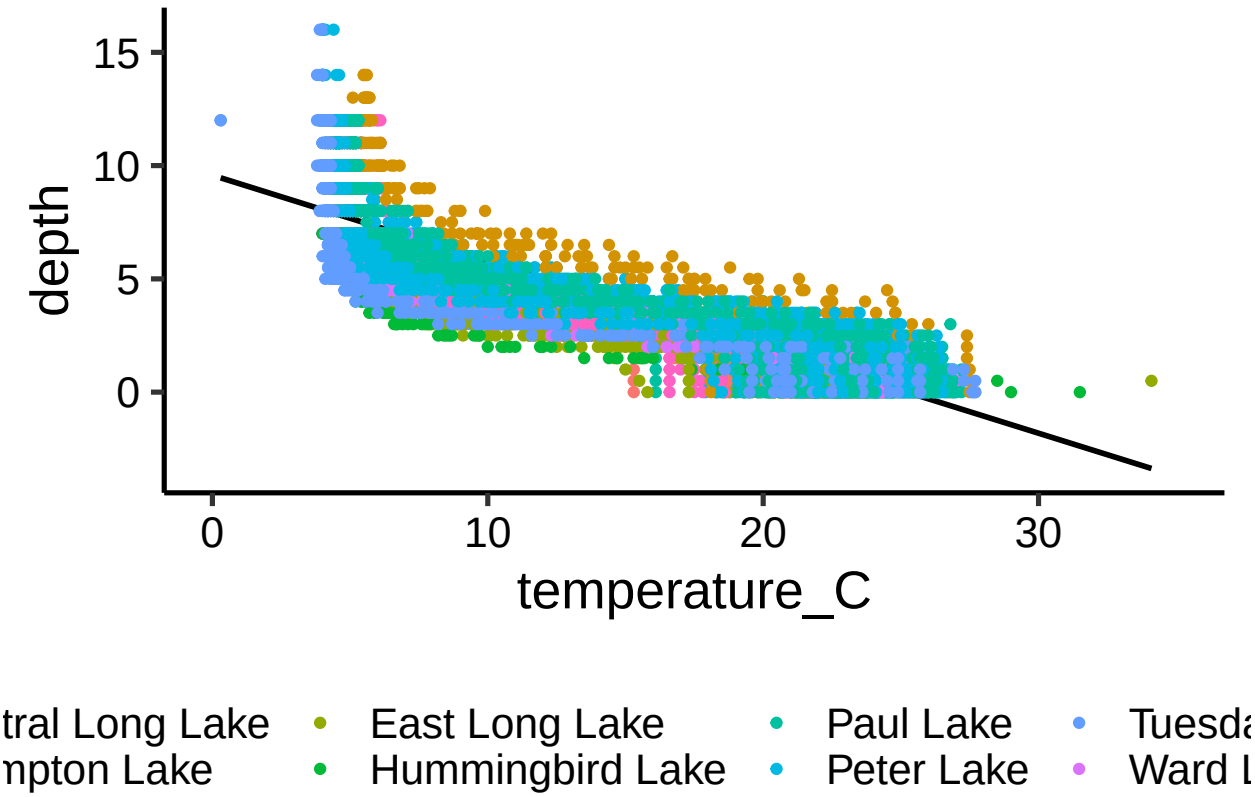
```
#4
LakeChemPhys.subset<-
  LakeChemPhys %>%
  select(lakename, sampledate, year4, daynum, depth, temperature_C) %>%
  na.omit()

LakeChemPhys.subset<-mutate(LakeChemPhys.subset,month=month(sampledate)) %>%
  filter(month=="7")
```

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.

```
#5
tempbydepth <-
  ggplot(LakeChemPhys.subset, aes(x = temperature_C, y = depth, color=lakename)) +
  xlim(0,35)+
  geom_smooth(method = "lm", color="black") +
  geom_point()
print(tempbydepth)
```

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



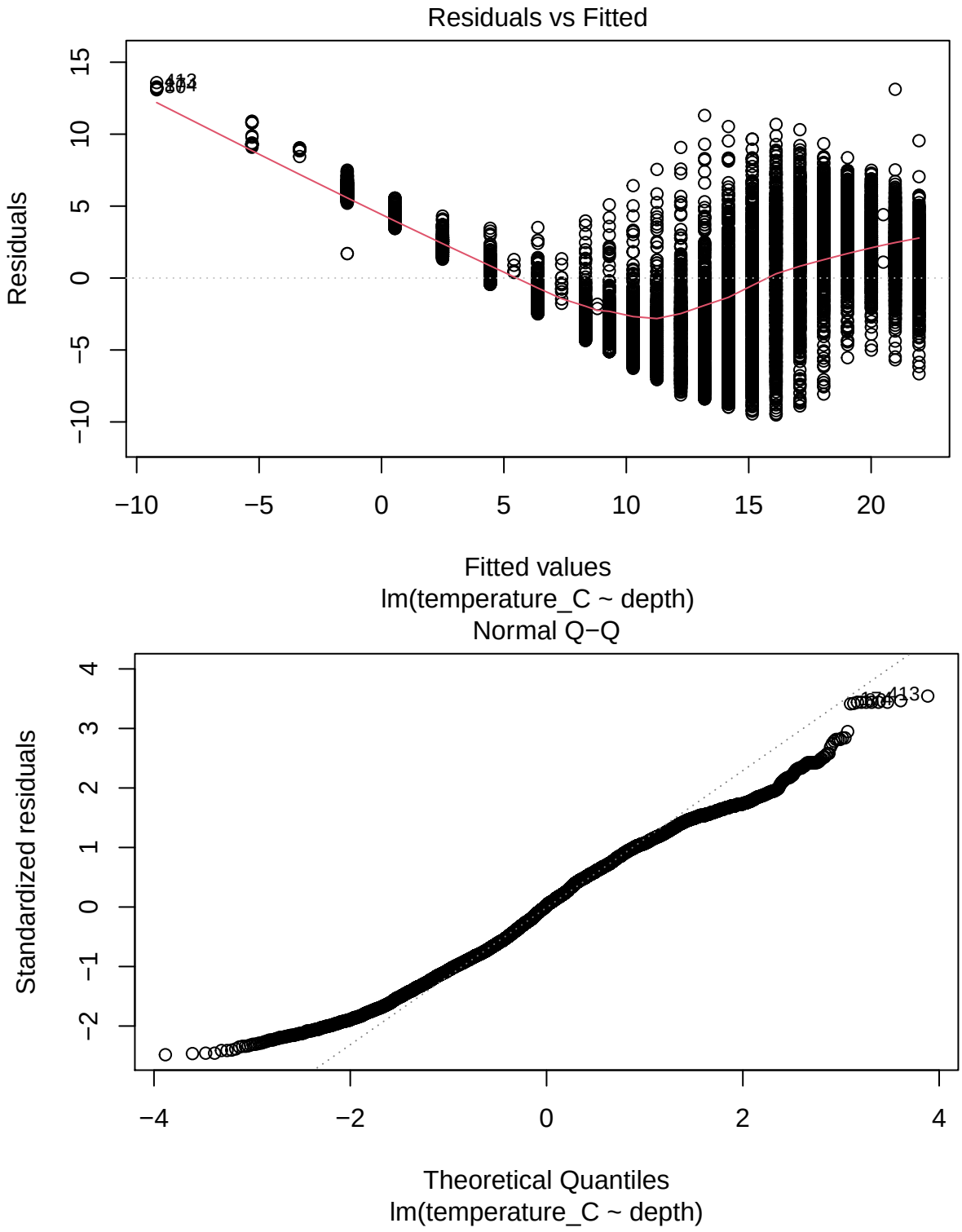
6. Interpret the figure. What does it suggest with regards to the response of temperature to depth? Do the distribution of points suggest about anything about the linearity of this trend?

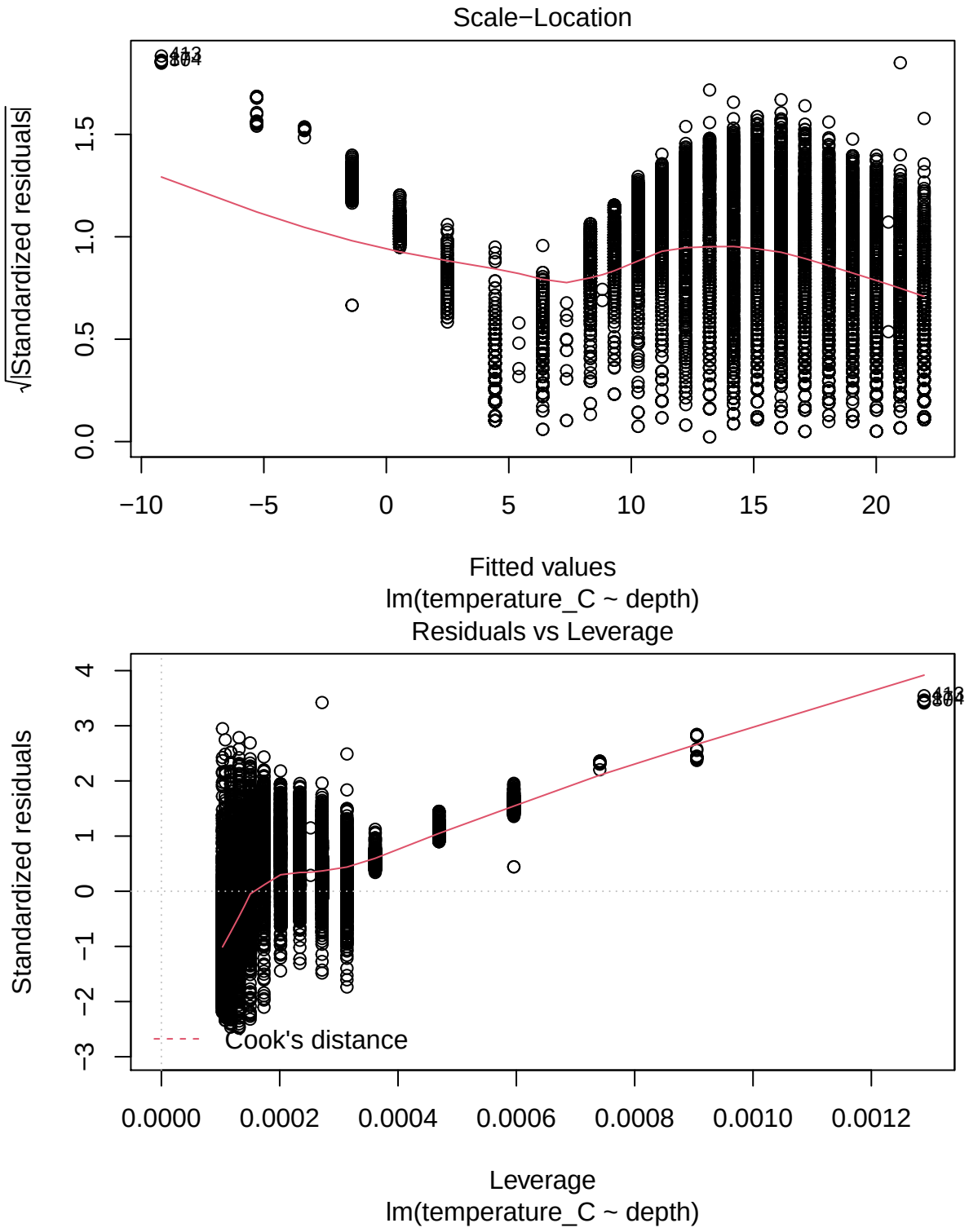
   Answer: As the water gets shallower, or less deep, the temperature of the water in the lake increases. The points suggest there is a linear relationship, but an exponential curve would probably better explain this.

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

```
#7
lmtempbydepth<- lm(data = LakeChemPhys.subset, temperature_C ~ depth)

plot(lmtempbydepth)
```


Residuals vs Fitted
Residuals
Fitted values
lm(temperature_C ~ depth)
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(temperature_C ~ depth)




```
summary(lmtempbydepth)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = LakeChemPhys.subset)
##
```

```
## 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
```

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 suggests that depth does not fully explain the changes in variability in temperature. The model explains 73.87% of the variability. As depth increases, there is a 1.94 degree Celsius decrease in temperature.

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
LCPAIC <- lm(data = LakeChemPhys.subset, temperature_C ~ year4 + daynum + depth)
summary(LCPAIC)
```

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

```
step(LCPAIC)
```

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

```
#10
LCPmodel <- lm(data = LakeChemPhys.subset, temperature_C ~ depth +daynum +year4)
summary(LCPmodel)
```

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

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 variability in the model, a small (~.2) improvement in the 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

#Simple Linear Regression
LCP.regression.1 <- lm(LakeChemPhys.subset$month ~ LakeChemPhys.subset$temperature_C)
print(LCP.regression.1)
```

```
##
## Call:
## lm(formula = LakeChemPhys.subset$month ~ LakeChemPhys.subset$temperature_C)
##
## Coefficients:
##                         (Intercept)  LakeChemPhys.subset$temperature_C
##                           7.000e+00                           7.048e-16
```

```
summary(LCP.regression.1)
```

```
##
## Call:
## lm(formula = LakeChemPhys.subset$month ~ LakeChemPhys.subset$temperature_C)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.900e-14 -1.000e-14 -2.000e-15  1.000e-15  3.827e-11
##
## Coefficients:
##                                   Estimate Std. Error   t value Pr(>|t|)
## (Intercept)                      7.000e+00  7.747e-15 9.036e+14   <2e-16 ***
## LakeChemPhys.subset$temperature_C 7.048e-16  5.246e-16 1.344e+00    0.179
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.881e-13 on 9726 degrees of freedom
## Multiple R-squared:    0.5,  Adjusted R-squared:  0.4999
## F-statistic:  9726 on 1 and 9726 DF,  p-value: < 2.2e-16
```

```
cor.test(LakeChemPhys.subset$month, LakeChemPhys.subset$temperature_C)
```

```
## Warning in cor(x, y): the standard deviation is zero
```

```
##
##  Pearson's product-moment correlation
##
## data:  LakeChemPhys.subset$month and LakeChemPhys.subset$temperature_C
## t = NA, df = 9726, p-value = NA
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  NA NA
## sample estimates:
## cor
##  NA
```

```r
#Multiple Linear Regression

LCP.regression.2 <- lm(data = LakeChemPhys.subset,
                  month ~ temperature_C + depth)
summary(LCP.regression.2)
```

```
##
## Call:
## lm(formula = month ~ temperature_C + depth, data = LakeChemPhys.subset)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.400e-14 -9.000e-15 -4.000e-15  0.000e+00  3.827e-11
##
## Coefficients:
##                 Estimate Std. Error    t value Pr(>|t|)
## (Intercept)    7.000e+00  2.356e-14  2.971e+14   <2e-16 ***
## temperature_C  2.259e-16  1.026e-15  2.200e-01    0.826
## depth         -1.262e-15  2.324e-15 -5.430e-01    0.587
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.881e-13 on 9725 degrees of freedom
## Multiple R-squared:    0.5,  Adjusted R-squared:  0.4999
## F-statistic:  4863 on 2 and 9725 DF,  p-value: < 2.2e-16
```

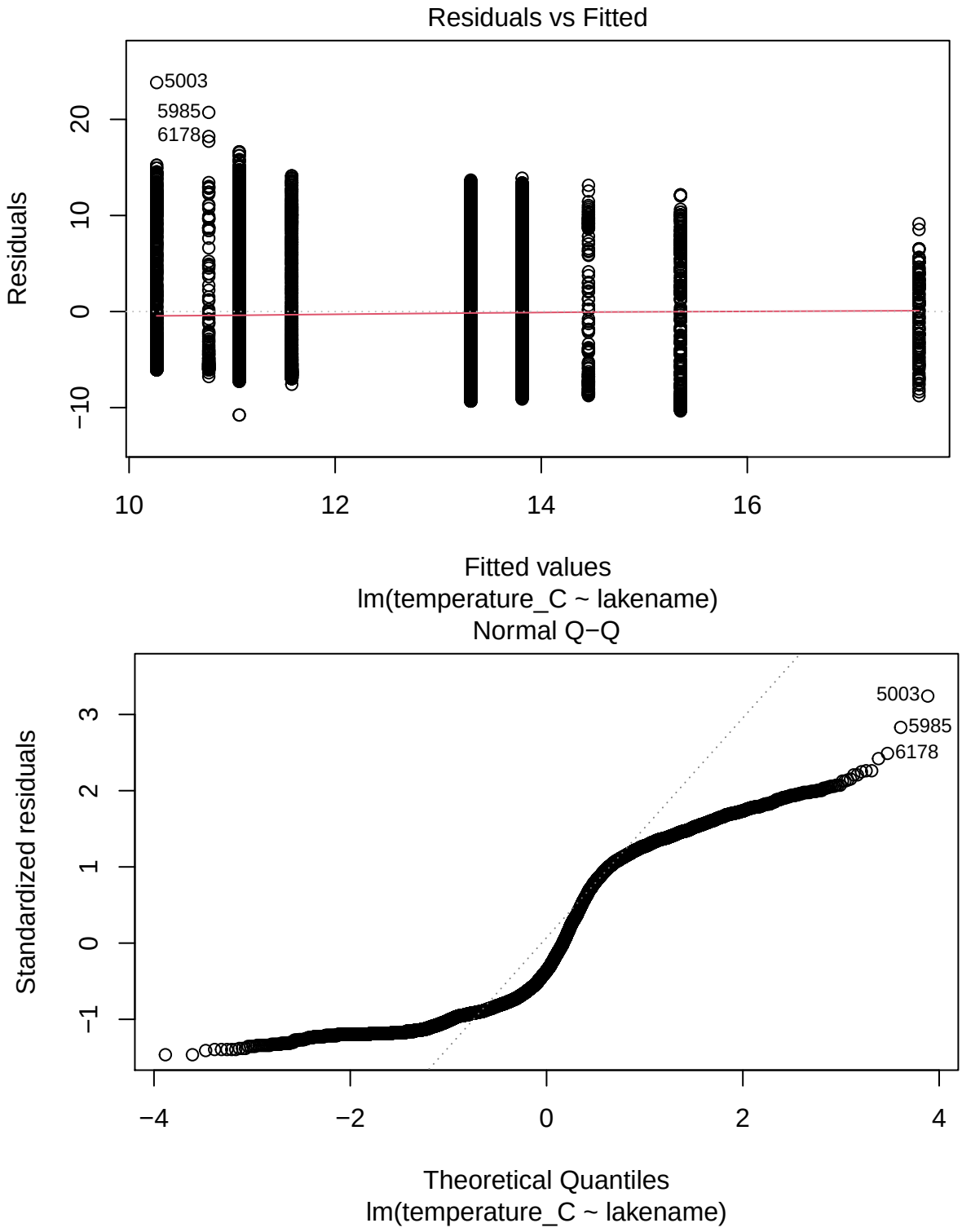
```r
#Anova
LakeChemPhys.subset.anova <- aov(data = LakeChemPhys.subset, temperature_C ~ lakename)
summary(LakeChemPhys.subset.anova)
```

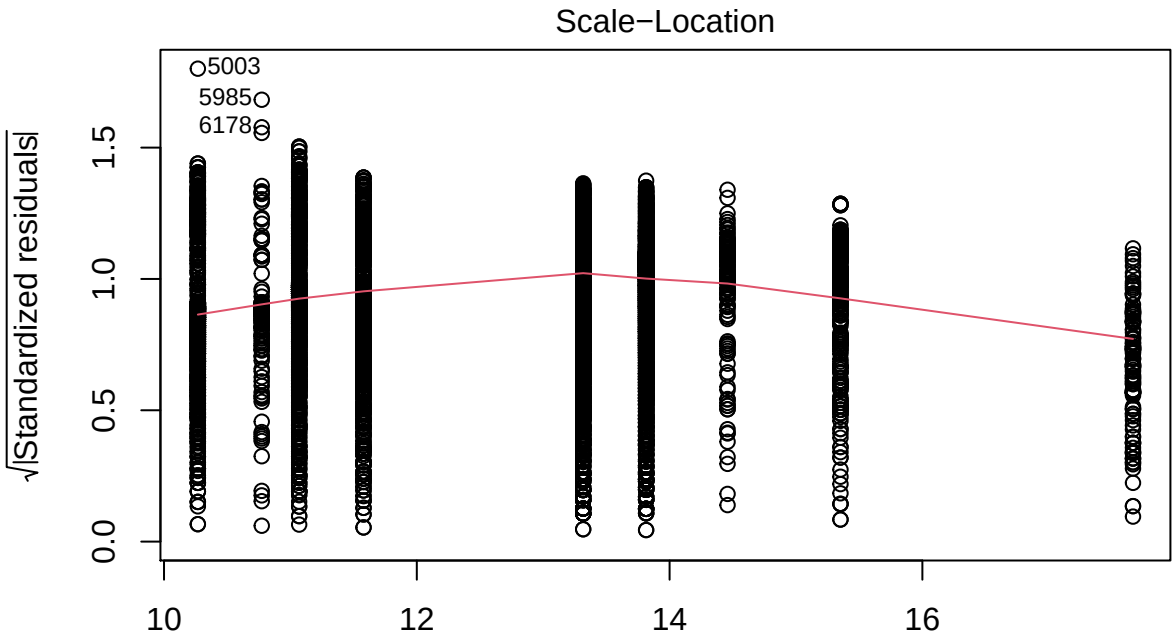
```
##               Df Sum Sq Mean Sq F value Pr(>F)
## lakename       8  21642  2705.2      50 <2e-16 ***
## Residuals   9719 525813    54.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

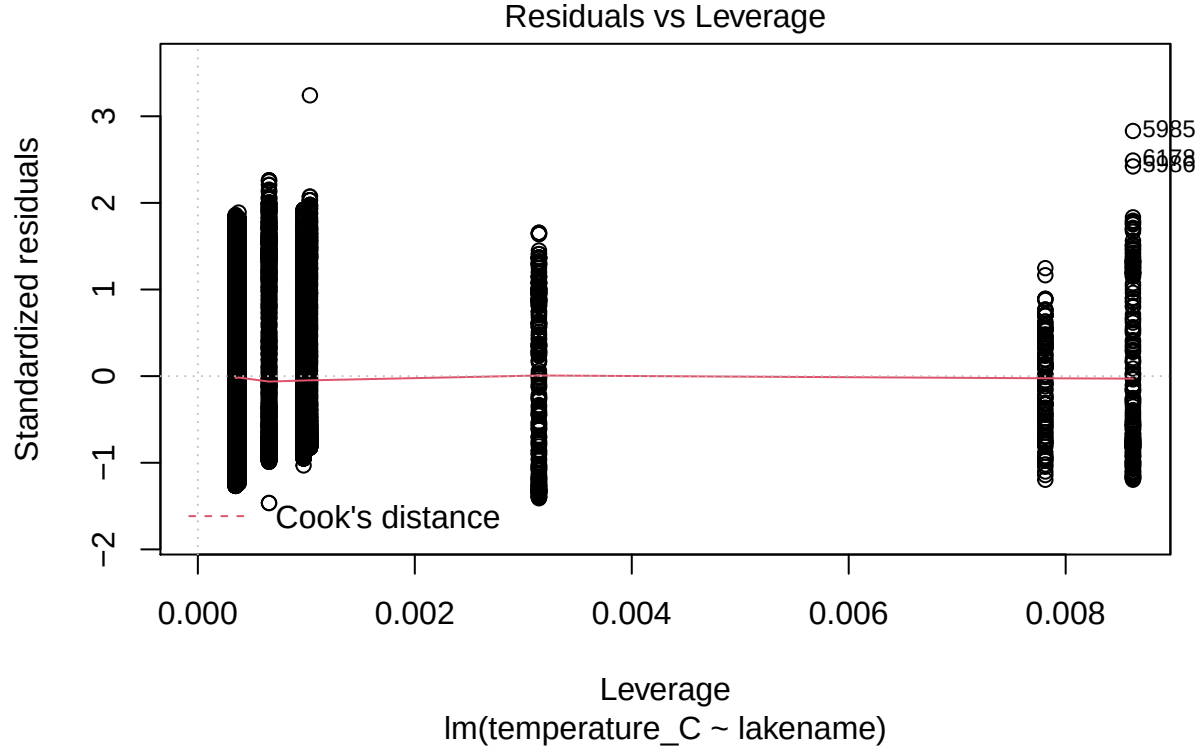
```r
LakeChemPhys.subset.anova2 <- lm(data = LakeChemPhys.subset, temperature_C ~ lakename)
summary(LakeChemPhys.subset.anova)
```

```
##               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
plot(LakeChemPhys.subset.anova2)
```


Residuals vs Fitted
5003
5985
6178
Residuals
20
10
0
−10
10
12
14
16
Fitted values
lm(temperature_C ~ lakename)
Normal Q−Q
5003
5985
6178
Standardized residuals
3
2
1
0
−1
−4
−2
0
2
4
Theoretical Quantiles
lm(temperature_C ~ lakename)





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

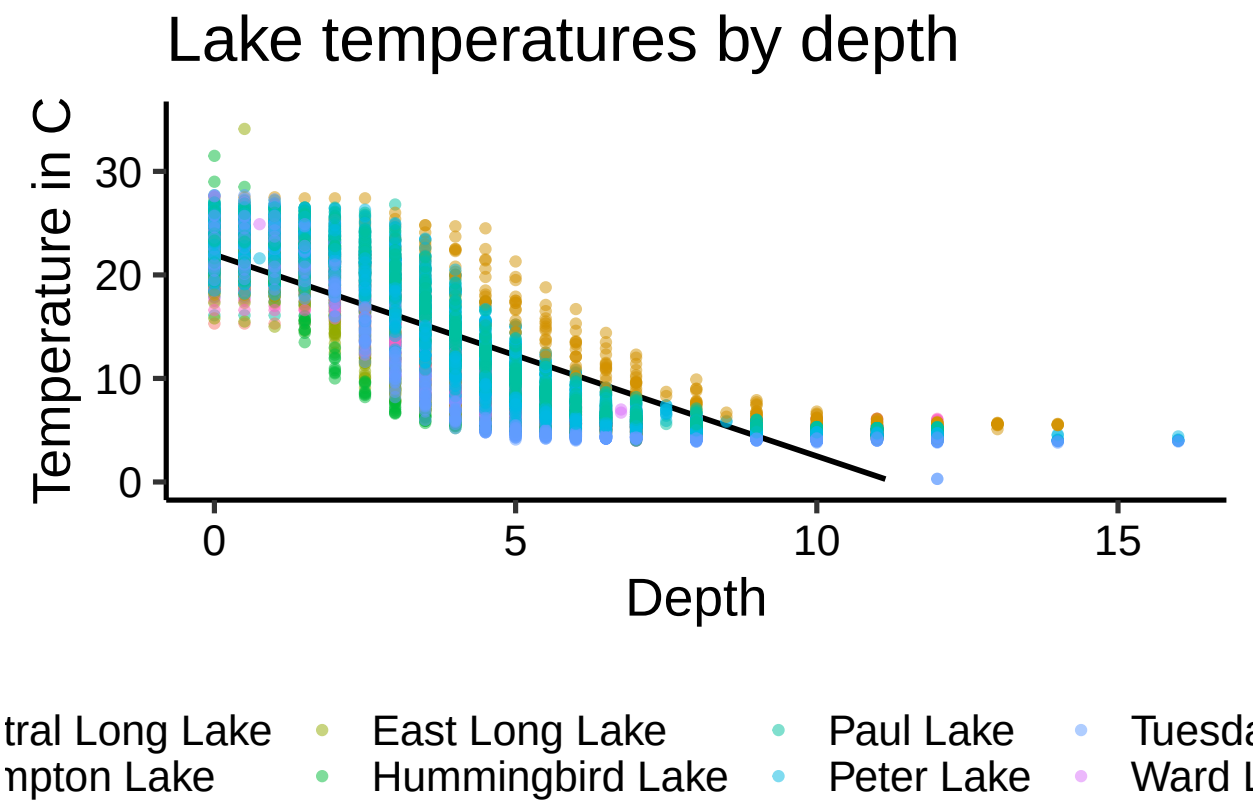
    Answer: There is a statistical difference in the means because the p values are stastically different <.05. The linear model explains a tiny bit of variability in the model, p value is less than alpha, so we can reject the null hypothesis. The multiple linear regression model explains 50% of the variability.

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.
tempbydepth2 <-
  ggplot(LakeChemPhys.subset, aes(y = temperature_C, x = depth, color=lakename)) +
  ylim(0,35)+
  geom_smooth(method = "lm", se=FALSE, color="black") +
  labs(lakename= "Lake Name", title="Lake temperatures by depth", x="Depth", y="Temperature in C")+
  geom_point(alpha=.5)
print(tempbydepth2)
```

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

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



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

```
#15

TukeyHSD(LakeChemPhys.subset.anova)
```

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

```
##LakeChemPhys.groups<- HSD.test(LakeChemPhys.subset.anova, "lakename", group = TRUE)
##print(LakeChemPhys.groups)
```

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?

If higher than 0.05 they are statistically the same **Only one with statistically the same is Peter/Paul Lake**

> Answer: Below I categorized the lakes. If they had a p adjusted value of less than or equal to 0.05 they were statistically different, if they had a p adjusted value of more than 0.05. No lake has a mean temperature statistically distinct form all other lakes.

#Smaller than 0.05 Paul Lake - Central Long Lake Ward Lake - Central Long Lake Hummingbird Lake-Crampton Lake Paul Lake-Crampton Lake Peter Lake-Crampton Lake
Peter Lake-Hummingbird Lake Ward Lake-Hummingbird Lake
Ward Lake-Peter Lake
Ward Lake-Paul Lake
West Long Lake-Ward Lake

#0 - less than 0.05 East Long Lake - Central Long Lake Hummingbird Lake - Central Long Lake Peter

Lake - Central Long Lake Tuesday Lake - Central Long Lake West Long Lake-Central Long Lake East Long Lake-Crampton Lake Tuesday Lake-Crampton Lake West Long Lake-Crampton Lake
Paul Lake-East Long Lake
Peter Lake-East Long Lake Tuesday Lake-Paul Lake West Long Lake-Paul Lake Tuesday Lake-Peter Lake West Long Lake-Peter Lake Ward Lake-Tuesday Lake

#Greater than 0.05 Crampton Lake - Central Lake Ward Lake-Crampton Lake Hummingbird Lake-East Long Lake Tuesday Lake-Hummingbird Lake West Long Lake-Hummingbird Lake Peter Lake-Paul Lake West Long Lake-Tuesday 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: I believe we could explore this through a Welch's test is a two sample t test thattests if two populations, or in this case, lakes, have equal means.