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

Danlei Zou

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

# checking working directory
getwd()
```

```
## [1] "/Users/danleizou/EDA-Fall2022"
```

```
# loading necessary packages
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   1.2.1      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(viridis)
```

```
## Loading required package: viridisLite
```

```
library(lubridate)
```

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

```
# loading data set
LakeChemPhys <- read.csv("./Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv", stringsAsFactors = TRUE)

# setting sampledate column to date objects
LakeChemPhys$sampledate <- as.Date(LakeChemPhys$sampledate, format = "%m/%d/%y")

# checking sampledate column
class(LakeChemPhys$sampledate)
```

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

```
# 2

# Set theme
mytheme <- theme_classic(base_size = 14) + theme(axis.text = element_text(color = "black"),
    legend.position = "top")
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)

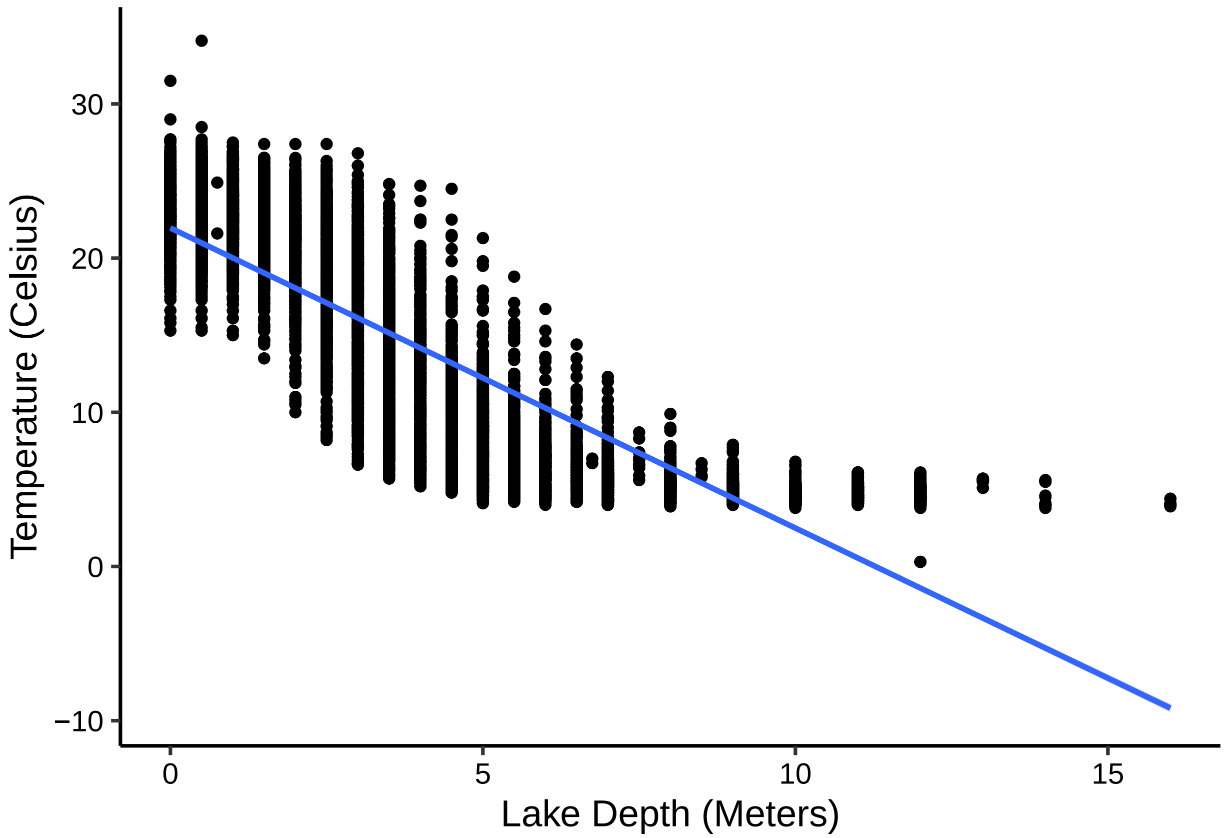
5. Visualize the relationship among the two continuous variables with a scatter plot of temperature by depth. Add a smoothed line showing the linear model, and limit temperature values from 0 to 35 °C. Make this plot look pretty and easy to read.

```r
# 4

# wrangle dataset according to criteria
LakeChemPhys.processed <- LakeChemPhys %>%
    filter(month(sampledate) == 7) %>%
    select(lakename, year4, daynum, depth, temperature_C) %>%
    drop_na(lakename, year4, daynum, depth, temperature_C)

# 5

# scatterplot to visualize relationship between temp and depth
LakeChemPhys.tempdepth <- ggplot(LakeChemPhys.processed, aes(x = depth, y = temperature_C)) +
    geom_point() + geom_smooth(method = lm, formula = y ~ x) + labs(x = "Lake Depth (Meters)",
    y = "Temperature (Celsius)")
print(LakeChemPhys.tempdepth)
```



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: According to the scatterplot, it does indicate a correlation between lake depth and temperature. The plot shows that temperature decreases as lake depth decreases. The points do show a relatively strong correlation of this, especially early on as depth decreases.

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

```r
# 7

# generating linear regression for LakeChemPhys.processed
LakeChemPhys.lm <- lm(data = LakeChemPhys.processed, temperature_C ~ depth)
summary(LakeChemPhys.lm)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = LakeChemPhys.processed)
##
## 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 Adjusted R-squared is 0.73876, so nearly 74% of the variability in temperature is based on changes in depth, and this finding is based on 9276 degrees of freedom. The result is statistically significant because our p-value is $< 2.2\text{e-}16$, which is significantly different from 0 and also indicates that depth is a statistically significant indicator of lake temperature in July. For every 1m change in depth, temperature changes by nearly 22 degrees Celsius (Estimate Std. of 21.95597).

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

# AIC to determine best explanatory variables for temp
LakeChemPhys.AIC <- lm(data = LakeChemPhys.processed, temperature_C ~ depth + year4 +
    daynum)

# choosing model by AIC in Stepwise Algorithm
step(LakeChemPhys.AIC)
```

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

```r
# 10

# running multiple regression on recommended set of variables
LakeChemPhys.multiple.lm <- lm(formula = temperature_C ~ depth + year4 + daynum,
    data = LakeChemPhys.processed)
summary(LakeChemPhys.multiple.lm)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = LakeChemPhys.processed)
##
## 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 ***
## year4        0.011345   0.004299    2.639  0.00833 **
## daynum       0.039780   0.004317    9.215  < 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 final set of explanatory variables the AIC method suggests we use to predict temperature in the multiple regression are depth, year4, and daynum. This model's Adjusted R-squared is 0.7411, which isn't much different from our previous model in #7's Adjusted R-squared of 0.73876. This model with depth, year4, and daynum explain 74.11% of the observed variance which is very close to the previous model explaining 73.876% of the observed variance, so it can't be considered a significant improvement.

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 model for average temp for lakes in July
LakeChemPhys.anova.lakes <- aov(data = LakeChemPhys.processed, temperature_C ~ lakename)
summary(LakeChemPhys.anova.lakes)
```

```
##               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
```

```
# linear model for average temp for lakes in July
LakeChemPhys.lm.lakes <- lm(data = LakeChemPhys.processed, temperature_C ~ lakename)
summary(LakeChemPhys.lm.lakes)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = LakeChemPhys.processed)
##
## 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
```

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

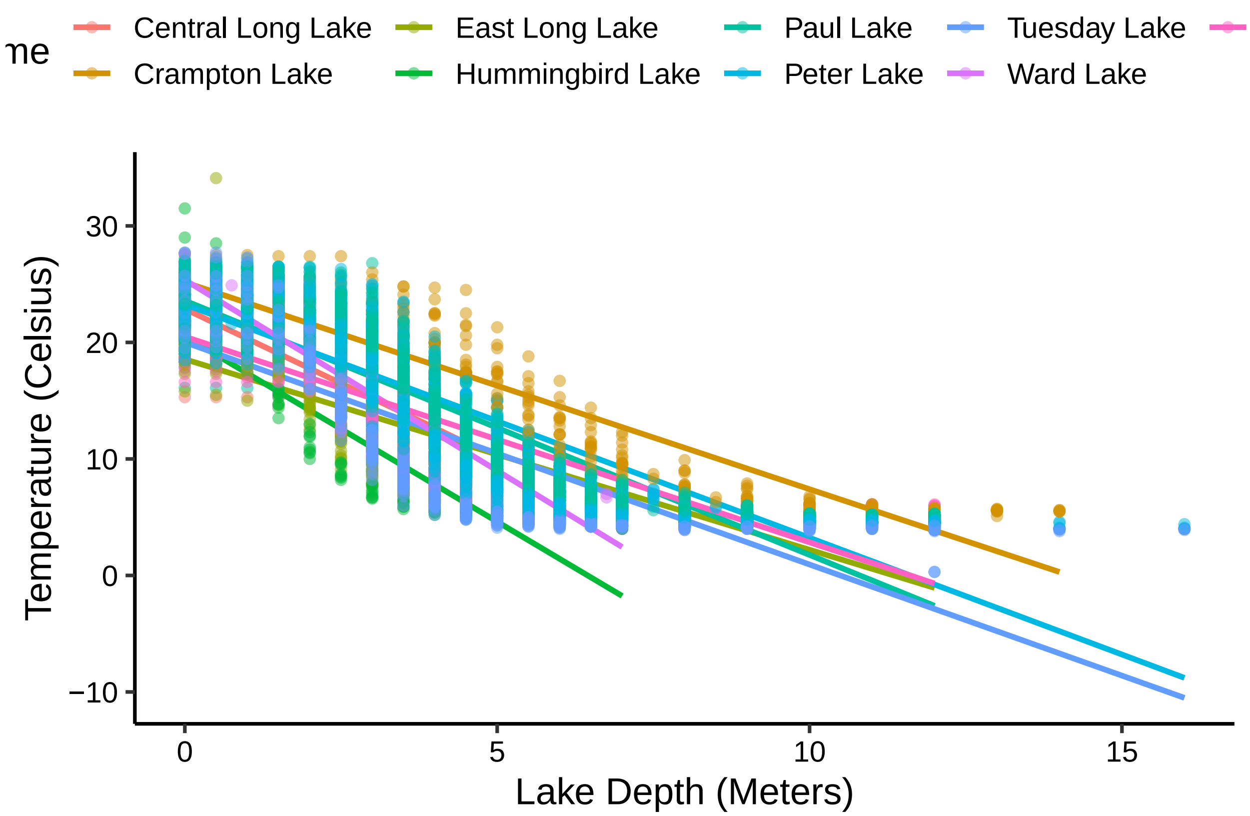
    Answer: The ANOVA model showed 8 degrees of freedom and a p-value of <2e-16, showing a significant difference in mean temperature among the different lakes. With the ANOVA model, we can now run post-hoc tests to determine which lakes are different. In the linear model we can see from the results that the different lakes all have different means, and are statistifcally significant since the p-value is <2.2e-16. However, the Adjusted R-squared is only 0.03874, so this linear model may not explain much of the variances in lake temperatures.

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 plot showing temp by depth for different lakes
ggplot(LakeChemPhys.processed, aes(x = depth, y = temperature_C, color = lakename)) +
    geom_smooth(method = "lm", se = FALSE) + geom_point(alpha = 0.5) + labs(x = "Lake Depth (Meters)",
    y = "Temperature (Celsius)")
```

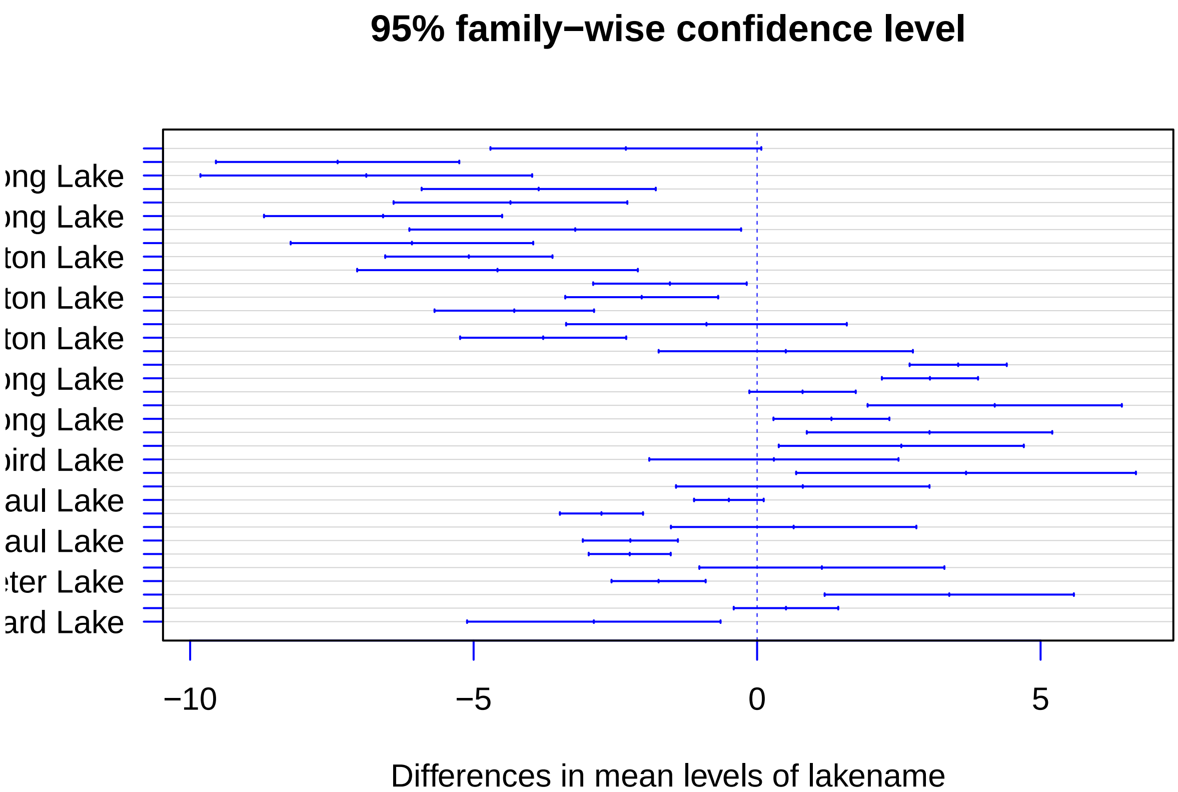
```
## `geom_smooth()` using formula 'y ~ x'
```



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

```
# 15

# using Tukey's HSD test to determine which lakes have different means
Tukey.lakes <- TukeyHSD(LakeChemPhys.anova.lakes)
plot(Tukey.lakes, las = 1, col = "blue")
```




```
# using Tukey's HSD test to determine groupings of pairwise relationships
Tukey.lakes.pairs <- HSD.test(LakeChemPhys.anova.lakes, "lakename", group = TRUE)
print(Tukey.lakes.pairs)
```

```
## $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"
```

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 has the same mean temperature as Peter Lake because they were both assigned the group "c". None of the lakes have a mean temperature that is statistically distinct from all the other lakes, because none of them were assigned a group that wasn't assigned to 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: The two-sample T-Test to see if Peter Lake and Paul Lake have distinct mean temperatures, because it would tell us if the means are equal.

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
# wrangle dataset according to criteria
Lakes.Crampton.Ward <- LakeChemPhys %>%
    filter(month(sampledate) == 7) %>%
    filter(lakename == "Crampton Lake" | lakename == "Ward Lake") %>%
    drop_na(lakename, temperature_C)

# running two-sample T-test on data to see if July temp is same or different
Lakes.Crampton.Ward.twosample <- t.test(Lakes.Crampton.Ward$temperature_C ~ Lakes.Crampton.Ward$lakenam
Lakes.Crampton.Ward.twosample
```

```
##
##  Welch Two Sample t-test
##
## data:  Lakes.Crampton.Ward$temperature_C by Lakes.Crampton.Ward$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 
```

```
## 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 two-sample T-test had 200 degrees of freedom and a p-value of 0.2649, and showed that the means temperatures for the lakes is not equal. The mean temperature of Crampton Lake is ~15.35 while the mean temperature of Ward Lake is ~14.46, which matches our answer for 16. In #16, the two are both shown to be b group but Crampton Lake is also in a group while Ward Lake was also in b group.