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

Lauren Shohan

Fall 2024

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
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## 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(corrplot)
```

```
## corrplot 0.94 loaded
```

```
library(agricolae)
library(lubridate)
options(scipen = 4)
library(here)
```

```
## here() starts at /home/guest/EDE_Fall2024
```

```
here()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```
getwd()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

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

Lake_Chem_Physics$sampledate <- mdy(Lake_Chem_Physics$sampledate)

#2
mytheme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "bottom",
        panel.border = element_rect(color = 'black', fill = NA, size = 0.8),
        panel.grid.minor = element_line(color = 'grey', size = 0.25)
        )
```

```
## Warning: The `size` argument of `element_rect()` is deprecated as of ggplot2 3.4.0.
## i Please use the `linewidth` argument instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

## Warning: The `size` argument of `element_line()` is deprecated as of ggplot2 3.4.0.
## i Please use the `linewidth` argument instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
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.

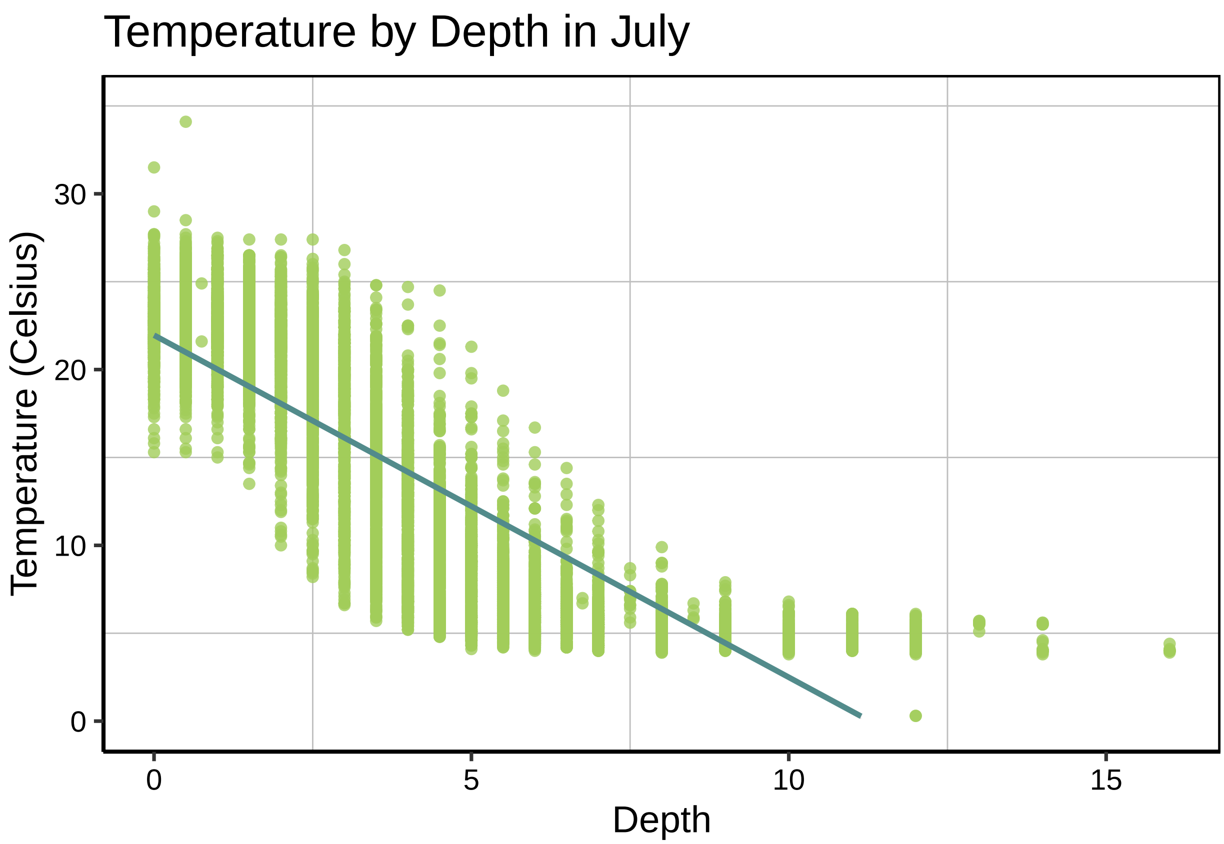
```
#4
Filtered_LakeData <- Lake_Chem_Physics %>%
  filter(month(sampledate) == 7) %>%
  select(lakename, year4, daynum, depth,temperature_C) %>%
  drop_na()


#5
temperaturebydepth <- Filtered_LakeData %>%
  ggplot(aes(x = depth, y = temperature_C)) +
  geom_point(color = 'darkolivegreen3', alpha = 0.8) +
  geom_smooth(method = 'lm', color = 'darkslategray4') +
  ylim(0, 35) +
  labs(y = 'Temperature (Celsius)', x = 'Depth', title =
         'Temperature by Depth in July')

print(temperaturebydepth)
```

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

## Warning: Removed 24 rows containing missing values or values outside the scale range
## (`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 figure suggests there may be an inverse relationship between temperature to depth, that while depth increases the temperature decreases. However, the distribution of points to the linear line seems quite broad/far away from the line which may indicate a non-linear relationship or other factors influencing tempearture.

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

```
#7
temperature_regression <- lm(Filtered_LakeData$temperature_C
                             ~ Filtered_LakeData$depth) # y ~ x
summary(temperature_regression)
```

```
##
## Call:
## lm(formula = Filtered_LakeData$temperature_C ~ Filtered_LakeData$depth)
##
## 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 ***
## Filtered_LakeData$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: This linear regression shows a strong, statistically significant relationship between temperature and depth. The R-square value of 0.7387 explains that 73.87% of the variability in temperature is explained by changes in depth. This is based on 9,726 degrees of freedom which suggests a large sample size. The statistical significance between depth and temperature is very high with a pvalue <2.2e-16 which is an extremely low pvalue below the 0.05 threshold, which leads us towards the route of rejecting the null hypothesis. Based off the coefficienct estimate of depth being -1.946, temperature is predicted to change 1.946C every 1m change in depth.

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

Filtered_Lake_AIC <- lm(data = Filtered_LakeData, temperature_C ~
                          year4 + daynum + depth)
step(Filtered_Lake_AIC)
```

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

```
  #AIC didnt remove any variables thus all 3 variables are a good fit


#10
Lake_Multiple_Regression <- lm(data = Filtered_LakeData,
                               temperature_C ~ year4 + daynum + depth)
summary(Lake_Multiple_Regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = Filtered_LakeData)
##
## 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 final set of explanatory variables that the AIC suggests we use are all three explanatory variables: year4, daynum, and depth. As seen from the adjusted r-squared value, this multiple regression model explains 74.12% of the observed variance in temperature. The depth only model explained 73.87% of the variance, so thus this new model only represents 0.25% of an improvement which is very small 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).

```r
#12
lake_temperatures_anova <- aov(data = Filtered_LakeData,
                               temperature_C ~ lakename)
summary(lake_temperatures_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
lake_temperatures_anova2 <- lm(data = Filtered_LakeData,
                               temperature_C ~ lakename)
summary(lake_temperatures_anova2)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = Filtered_LakeData)
##
## 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: In my anova model, the fvalue is very high at 50 and pvalue is <2e-16 which strongly suggests there being a high statistical significant differences between mean temperatures in the lakes. The linear model provides similar information, that all the lakes have significantly different mean temperatures from the reference lake (intercept). The multiple rsquared shows that 3.95% of variance of temperature is explained by the differences in the lakes.

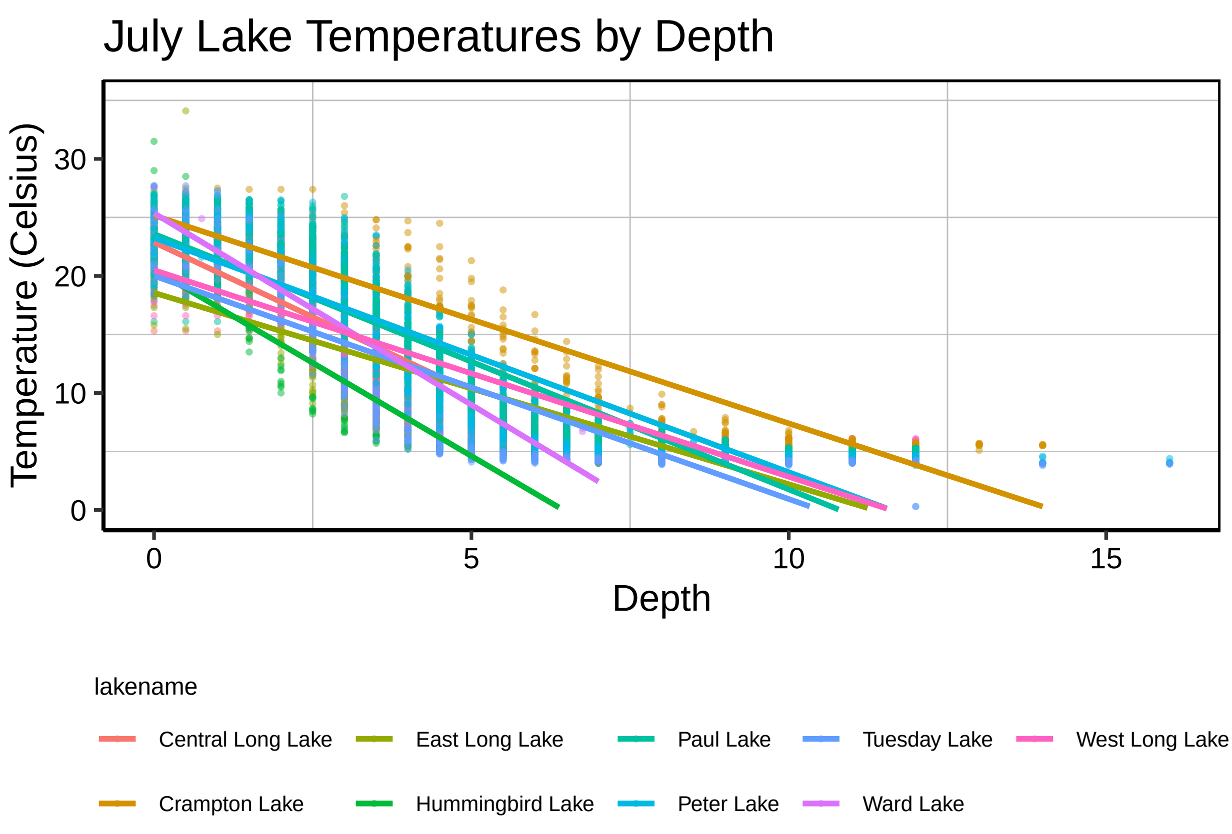
14. Create a graph that depicts temperature by depth, with a separate color for each lake. Add a geom_smooth (method = “lm”, se = FALSE) for each lake. Make your points 50 % transparent. Adjust your y axis limits to go from 0 to 35 degrees. Clean up your graph to make it pretty.

```
#14.
temperaturebydepthalllakes <- Filtered_LakeData %>%
  ggplot(aes(x = depth, y = temperature_C, color = lakename)) +
  geom_point(size = 0.6, alpha = 0.5) +
  geom_smooth(method = 'lm', se = FALSE) +
  ylim(0, 35) + #limits
  labs(y = 'Temperature (Celsius)', x = 'Depth',
       title = 'July Lake Temperatures by Depth') +
  theme(legend.text = element_text(size = 8),
        legend.title = element_text(size = 9),
        legend.title.position = 'top') #reformatting my legend


print(temperaturebydepthalllakes)
```

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

```
## Warning: Removed 73 rows containing missing values or values outside the scale range
## (‘geom_smooth()‘).
```



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

```r
#15
TukeyHSD(lake_temperatures_anova)
```

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

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: If the pvalue (padj) is more than 0.05, the difference between the lakes is not statistically significant. Comparing with Peter Lake, Paul Lake has a pvalue = 0.224 while Ward lake has a pvalue = 0.782, this indicates no statistical difference in mean temperatures between these

two lakes and Peter Lake. To find a lake that is statistically distinct from all other lakes, all its comparisons have pvalues <0.05. After looking through all the pvalues, none seem to have a mean temperature statistically distinct from 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: A two sample t-test would be a good next test to explore to see whether they have distinct mean temperatures. This would be more focused on comparing just these two lakes rather than the ANOVA which compared all the lakes. The t-test would provide a more detailed analysis comparison between these two lakes.

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
Crampton_Ward_Data <- Filtered_LakeData %>%
  filter(lakename == 'Crampton Lake' | lakename == 'Ward Lake')

Crampton_ward_lakes_ttest <- t.test(Crampton_Ward_Data$temperature_C
                                    ~ Crampton_Ward_Data$lakename)
Crampton_ward_lakes_ttest
```

```
##
##  Welch Two Sample t-test
##
## data:  Crampton_Ward_Data$temperature_C by Crampton_Ward_Data$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 test shows a t-stat of 1.1181, df = 200.37, and pvalue of 0.265. Since the pvalue is above 0.05, we fail to reject the null hypothesis and can say there isn't enough information to say Crampton and Ward Lakes mean temperatures are statistically different. Comparing with part 16, I said Ward Lake had no statistical difference in means with Peter Lake, thus saying Ward is similar to Peter Lake. In this ttest, we see Crampton Lake is similar to Ward Lake and so it is fair to say Crampton Lake might also be similar to Peter Lake which was not said outright part 16 results.