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

Gaby Czarniak

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.

```r
#1
# Check working directory and load packages
getwd()
```

```
## [1] "/home/guest/gaby-cz_EDE_Fall2023"
```

```r
library(tidyverse); library(lubridate)
library(htmltools)
library(dplyr)
library(cowplot); library(ggridges); library(ggthemes)
#install.packages("agricolae")
library(agricolae)
library(here)
here()
```

```
## [1] "/home/guest/gaby-cz_EDE_Fall2023"
```

```r
# Import data
Lake.chem.phys.raw <- read.csv(here(
  "Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv"),
  stringsAsFactors = TRUE)

# Set date columns to date format
Lake.chem.phys.raw$sampledate <- as.Date(
  Lake.chem.phys.raw$sampledate , format = "%m/%d/%y")
is.Date(Lake.chem.phys.raw$sampledate) #true
```

```
## [1] TRUE
```

```r
#2
# Build ggplot theme
gctheme <- theme_classic(base_size = 12) +
  theme(axis.text = element_text(color = "black"),
        # setting base format for legend
        legend.position = "right",
        legend.justification = "left",
        legend.title.align = 0)
# Set as default theme
theme_set(gctheme)
```

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 recorded during July does not change with depth across all lakes. (There is no relationship; slope is zero and intercept is zero.) Ha: Mean lake temperature recorded during 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.

```r
#4
# Wrangle the data
Lake.chem.phys.tempC <- Lake.chem.phys.raw %>%
  # including only dates in July
    # adding month column to select just July
    mutate(Month = month(sampledate)) %>%
    # Month
  filter(Month==7) %>%
  # only the columns we want
  select(
    lakename, year4, daynum,
    depth, temperature_C) %>%
  # only complete cases
  drop_na()
# summary(Lake.chem.phys.tempC)
glimpse(Lake.chem.phys.tempC)
```

```
## Rows: 9,728
## Columns: 5
## $ lakename      <fct> Paul Lake, Paul Lake, Paul Lake, Paul Lake, Paul Lake, P~
## $ year4         <int> 1984, 1984, 1984, 1984, 1984, 1984, 1984, 1984, 1984, 19~
## $ daynum        <int> 183, 183, 183, 183, 183, 183, 183, 183, 183, 183, 183, 1~
## $ depth         <dbl> 0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 5.0, 6.0, 7~
## $ temperature_C <dbl> 22.8, 22.9, 22.8, 22.7, 21.7, 20.3, 18.2, 14.8, 12.3, 8.~
```

```
#5
july_tempC_plot <-
  ggplot(Lake.chem.phys.tempC, aes(x=depth, y=temperature_C)) +
    geom_point() +
    # adjusting axes to hide extreme values
    xlim(0, 17) +
    ylim(0,35) +
    # finding a line of best fit
    geom_smooth(method = lm, color = "black") +
    ggtitle("Change in Temperature (Celsius) by Depth (meters)") +
    labs(x="Depth (m)", y="Temperature (C)")
print(july_tempC_plot)
```

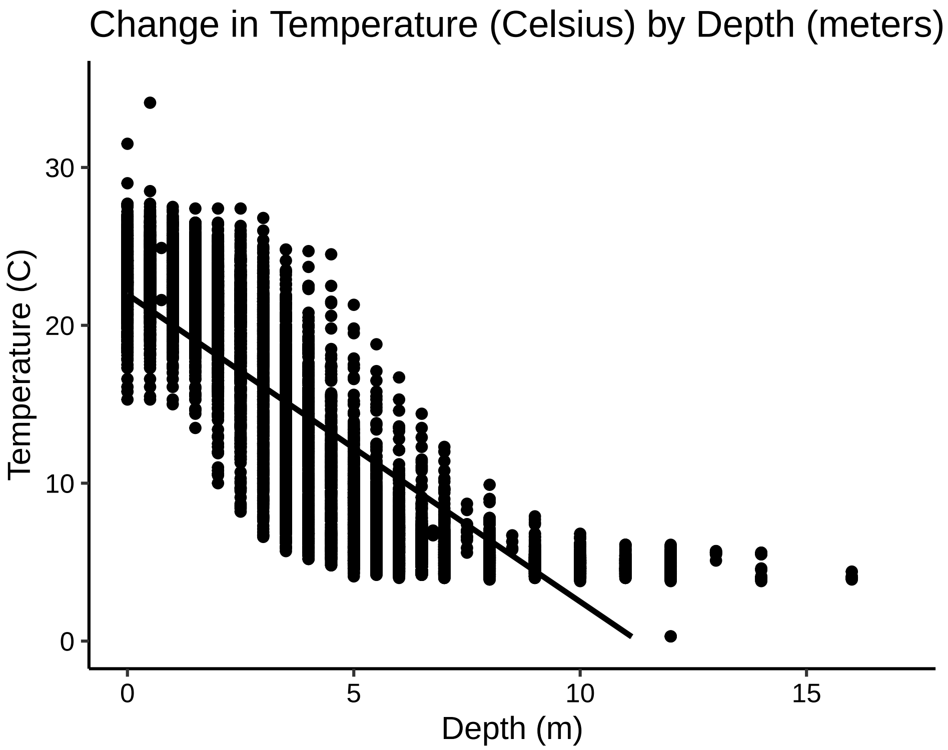


Figure 1: Change in Temperature (Celsius) by Depth (m)

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 that temperature decreases as depth increases. The distribution of points suggest that, at a certain point of depth in these lakes, the temperature evens out and stops getting much colder for the most part. So, beyond about 11m of depth, the line of best fit no longer does a good job at describing the relationship between temperature and depth.

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

```
#7
# simple linear regression providing lm with y and x
july_tempC_regression <- lm(data = Lake.chem.phys.tempC, temperature_C ~ depth)
summary(july_tempC_regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = Lake.chem.phys.tempC)
##
## 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
```

```
# correlation between the two variables
cor.test(Lake.chem.phys.tempC$temperature_C, Lake.chem.phys.tempC$depth)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Lake.chem.phys.tempC$temperature_C and Lake.chem.phys.tempC$depth
## t = -165.83, df = 9726, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8646036 -0.8542169
## sample estimates:
##        cor
## -0.8594989
```

```
# -0.86 signifies a negative correlation and
# a strong correlation between the two variables

# plotting the regression
# par(mfrow = c(2,2), mar = c(4,4,4,4))
# plot(july_tempC_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 residuals, or errors, will range from -9.52 to 13.58 and the median is 0.06. The R-squared value of 0.7387 signifies that depth is explaining around 74% of the variability in temperature. These findings are based on 9726 degrees of freedom, which are based on the number of observations in the sample and the number of variables being considered in the linear regression. The p-value is < 0.05, which is the confidence level. This means that the coefficient (related to the correlation among the variables) is statistically different than zero so it is, indeed, worthwhile to try to estimate temperature in lakes based on depth. My intercept term is 21.96 and the slope of my regression line that is trying to find the relationship between depth and temperature is -1.95 (the predicted temperature decrease, in degrees Celsius, per every 1m increase in depth), so there is a negative relationship; as the lake depth increases, the temperature decreases. This is different from the null hypothesis, which assumes an intercept and slope of 0. Since all p-values proved smaller than the confidence level, we can conclude that the regression was meaningful–we can, indeed, explain temperature by 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
# Taking note of initial AIC for comparison
TemperatureAIC_justdepth <- lm(data = Lake.chem.phys.tempC,
                     temperature_C ~ depth)
# step(TemperatureAIC_justdepth) -- commenting out for length purposes
# just depth: AIC is 26153.25

# Running Akaike's Information Criterion (AIC)
TemperatureAIC <- lm(data = Lake.chem.phys.tempC,
                     temperature_C ~ year4 +
                     daynum + depth)
# TemperatureAIC
# Choose a model by AIC in a Stepwise Algorithm
step(TemperatureAIC)
```

```
## 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 = Lake.chem.phys.tempC)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##    -8.57556      0.01134      0.03978     -1.94644
```

```
# more explanatory variables: AIC is 26065.53
# AIC decreases when we add the explanatory variables

#10
Temperature_model <- lm(data = Lake.chem.phys.tempC,
                        temperature_C ~ year4 +
                        daynum + depth)
summary(Temperature_model)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = Lake.chem.phys.tempC)
##
## 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
```

```
# The new R-squared is .7412 so
# temperature is slightly better explained
# by this new, multiple regression.
```

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 suggests we use year4, daynum, and depth to predict temperature in our multiple regression–no explanatory variables were returned over the “” line. The model with only depth as the explanatory variable has an AIC of 26153.25, whereas the model with daynum and year4 as additional explanatory variables has an AIC of 26065.53, so adding the explanatory variables decreases the AIC, and we know a lower AIC is better. This expanded model explains 74.12% of our observed variance, which is ever so slightly better than the 73.87% I had rounded up to 74% in the model using only depth. So this model explains .25% more observed variance than the only-depth model. The improvement over the model that uses only depth as the explanatory variable is there, but it’s minimal–removing year4 or daynum would both increase the AIC, which is the opposite of what we want, so we want to keep them in.

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
# null = the averages are equal
# alt = at least one pair of means is not equal
# factor = the lake temperatures
# levels = the lakes
# assume balanced design

# summary(Lake.chem.phys.tempC$temperature_C)
# summary(Lake.chem.phys.tempC)
# mean has high variance if compared with min and max values
# want to understand if this is due to the difference in lake
# summary(Lake.chem.phys.tempC$lakename) -- commenting out for length purposes
# note: not a balanced experiment because
# we have more samples at some lakes than others

# Format ANOVA as aov
# specify continuous, dependent variable and categorical variable
TempC_totals_anova <- aov(data = Lake.chem.phys.tempC,
                          temperature_C ~ lakename)
summary(TempC_totals_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
```

```
# results of the AOV
# 8 degrees of freedom
# deviation of each obs from the mean = 21642
# p value is < 0.05 so we reject the null hypothesis that
# the averages are equal across lakes
# plot(TempC_totals_anova) -- commenting out for length purposes

# Format ANOVA as lm
TempC_totals_anova_lm <- lm(data = Lake.chem.phys.tempC,
                            temperature_C ~ lakename)
summary(TempC_totals_anova_lm)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = Lake.chem.phys.tempC)
##
## 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
```

```
# main diff btwn lm and aov is the output and summary table
# summary from AOV obj is a traditional ANOVA output
# summary for linear regression will have one row for each level
# plot(TempC_totals_anova_lm) -- commenting out for length purposes
```

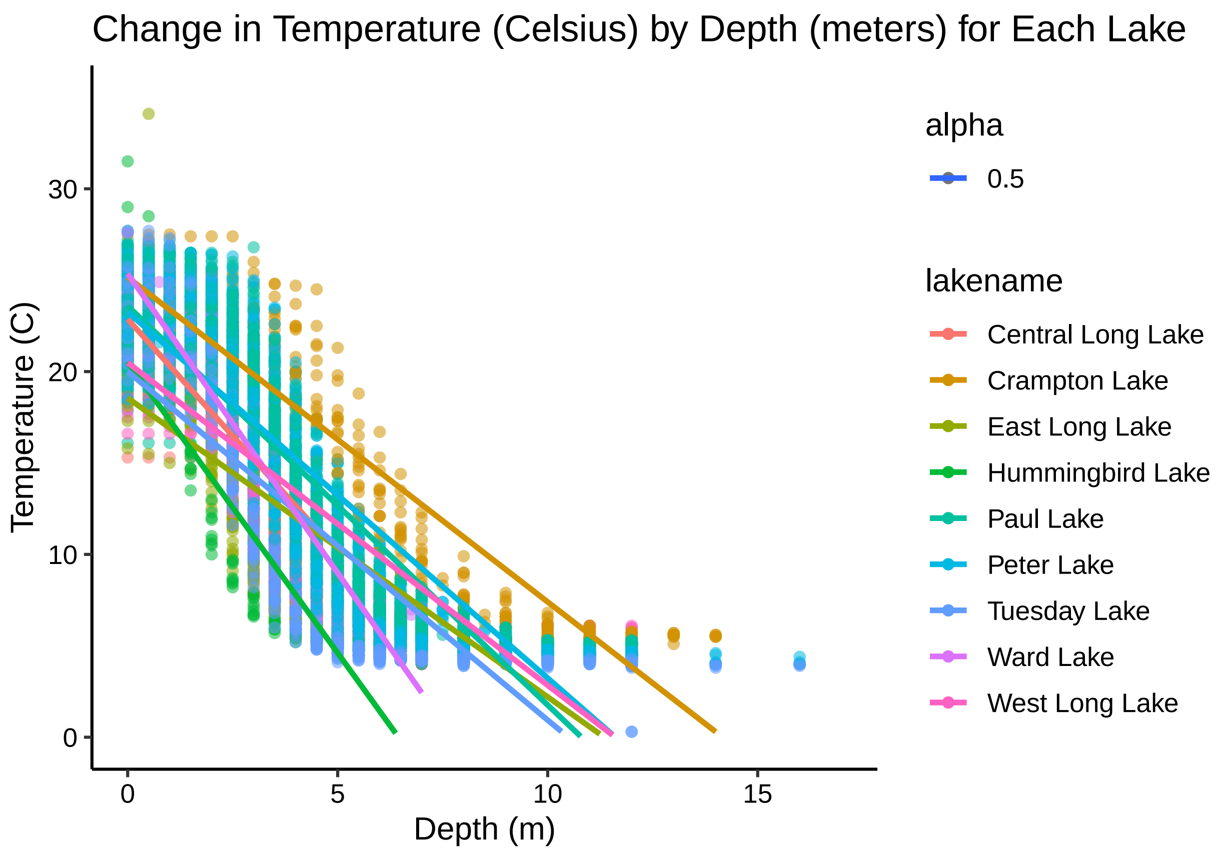
13. Is there a significant difference in mean temperature among the lakes? Report your findings.

    Answer: Yes, there are significant differences in mean temperature among the lakes. The means are not all the same across the different sites. In the "Estimate" column of calling the summary on the ANOVA as lm, the averages for each lake differ from the base (Intercept, or Central Long Lake). The mean temperature of Central Long Lake is 17.67 degrees Celsius, while that of Crampton Lake is 15.53, that of East Long Lake is 10.27, and so on. The p-value is <.05 again, so we can reject the null hypothesis.

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.
july_tempC_plot_2 <-
  ggplot(Lake.chem.phys.tempC, aes(
    x=depth, y=temperature_C, color = lakename, alpha = 0.5)) +
    geom_point() +
    # adjusting axes to hide extreme values
    xlim(0, 17) +
    ylim(0,35) +
    # finding a line of best fit
    #geom_density(alpha = 0.5) +
    geom_smooth(method = "lm", se = FALSE) +
    ggtitle("Change in Temperature (Celsius) by Depth (meters) for Each Lake") +
    labs(x="Depth (m)", y="Temperature (C)")
print(july_tempC_plot_2)
```

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



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

```
#15
# Post-hoc test to determine which particular differences
# between pairs of means are significant
# Pair-wise comparison
# TukeyHSD(TempC_totals_anova) -- commenting out for length purposes
# Run HSD test
Temperature_differences_groups <-
  HSD.test(TempC_totals_anova,
           "lakename", group = TRUE)
Temperature_differences_groups
```

```
## $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: In terms of differences in means that are statistically significant, when I run the HSD test on the aov, I see that Central Long Lake and Crampton Lake have similar means, Paul Lake and Peter Lake have similar means, Tuesday Lake and Hummingbird Lake have similar means (not an exhaustive list, just examples). No one lake is entirely statistically distinct from all the other lakes, each one sees overlap in the groups from the HSD test.

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 Lake and Paul Lake, we might explore the two-sample t-test, because it is used to test the hypothesis that the mean of two samples is equivalent. If proven that the mean of the two samples is not equivalent, we would know that they have distinct mean temperatures, assuming the variance of the two groups is equivalent.

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 data to include only records for
# Crampton and Ward lakes
CramptonWard_tempC <- Lake.chem.phys.tempC %>%
  filter(lakename %in% c("Crampton Lake",
          "Ward Lake"))
# CramptonWard_tempC

# Running two-sample T-test
# null hypothesis is that the two lakes' means are the same
# CramptonWard_tempC$temperature_C will be continuous dependent variable
# CramptonWard_tempC$lakename will be categorical variable with two levels
# (Crampton Lake and Ward Lake)
TemperatureC.twosample <-
  t.test(CramptonWard_tempC$temperature_C ~
         CramptonWard_tempC$lakename)
TemperatureC.twosample
```

```
##
##  Welch Two Sample t-test
##
## data:  CramptonWard_tempC$temperature_C by CramptonWard_tempC$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
```

```
# p-value is > .05, so we cannot reject the null hypothesis.
# We cannot conclude that the means are meaningfully different.
```

Answer: The two-sample t-test returns a p-value of 0.2649, which is greater than 0.05. This tells us that we cannot reject the null hypothesis; we cannot conclude that the means of Crampton Lake and Ward Lake's temperatures are meaningfully different. This makes sense based on the HSD test in Q16; while I did not call these two out in comparison specifically there, Crampton Lake and Ward Lake overlapped in group b, leading me to believe the difference between the Crampton-Ward pair was not significant.