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

Aditi Jackson

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.

```
#1
# checking working directory is "ENV872 Setup (local)"
getwd()
```

```
## [1] "/Users/aditijackson/ENV872 Setup (local)"
```

```
# installing pacakges
# install.packages(tidyverse)
# install.packages(agricolae)
# install.packages("ggplot2")
# install.packages("ggthemes")

# loading packages
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
```

```
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## 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 error
```

```r
library(agricolae)
library(lubridate)
library(ggplot2)
library(ggthemes)

# importing data
LakeData_Raw <- read.csv("Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv",
                         stringsAsFactors = TRUE)

# converting date column to date objects with lubridate
LakeData_Raw$sampledate <- mdy(LakeData_Raw$sampledate)

#2
# building theme
my_theme <- theme_base(base_size = 10)  + theme(
  axis.text = element_text(color = "black"),# setting axis text color
  plot.title = element_text(color = "black"),# setting title text color
  panel.background = element_rect(fill="lightyellow",color="blue"),
  legend.background = element_rect(color='grey',fill = 'white'),# setting legend color
  legend.position = 'bottom', # setting position of legend to bottom
  legend.justification = "center") # centering legend

# setting theme as default
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 does not change with depth (M == 0)
   Ha: Mean lake temperature changes with depth (M != 0)

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)

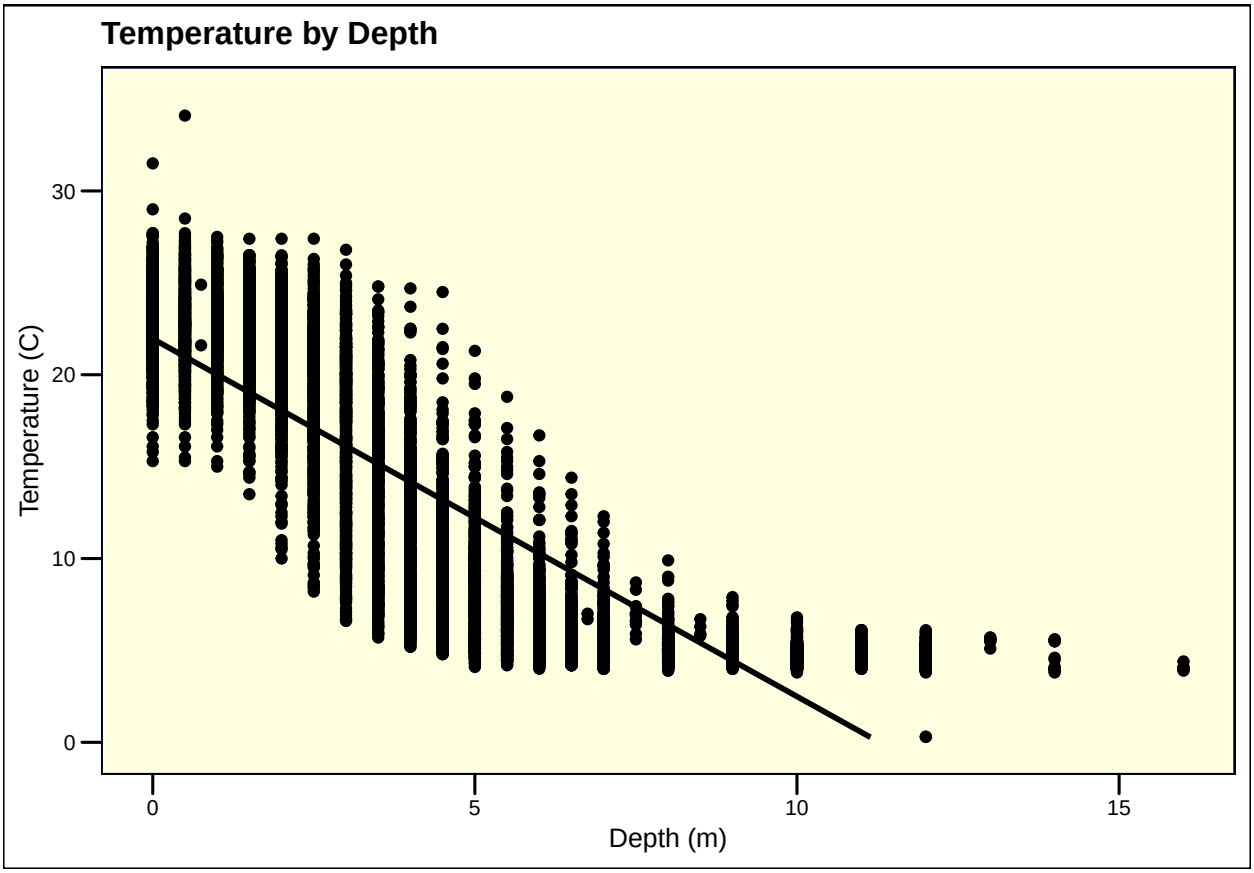
5. Visualize the relationship among the two continuous variables with a scatter plot of temperature by depth. Add a smoothed line showing the linear model, and limit temperature values from 0 to 35 °C. Make this plot look pretty and easy to read.

```r
#4
# adding a month column to the dataset
LakeData_Raw_Month <- mutate(LakeData_Raw, month = month(sampledate))

#wrangling data using a pipe function
LakeData_July <-
  LakeData_Raw_Month %>%
    # choosing rows with July
    filter(month==7) %>%
    # selecting columns: lakename, year4, daynum, depth, temperature_C
    select(lakename:daynum,depth,temperature_C) %>%
    # removing NAs
    drop_na()

#5
# creating scatter plot of temp by depth using ggplot
Lake_Scatter <- ggplot(LakeData_July,aes(x=depth,y=temperature_C))+
  geom_point()+
  geom_smooth(method=lm,SE=FALSE,color="black")+
  ylim(0,35)+
  labs(x="Depth (m)",y="Temperature (C)",title="Temperature by Depth")
```

```
## Warning in geom_smooth(method = lm, SE = FALSE, color = "black"): Ignoring
## unknown parameters: `SE`
```

```r
# displaying plot
Lake_Scatter
```

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

## Warning: Removed 24 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: There appears to be an inverse relationship between temperature and depth. For a given increase in depth, temperature decreases by some amount. The distribution of points suggests a negative linear relationship.

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

```
#7
# performing linear regression of temp and depth
Linear_Regression_Q7 <- lm(
  LakeData_July$temperature_C ~ LakeData_July$depth)

# displaying regression results
summary(Linear_Regression_Q7)
```

```
##
## Call:
## lm(formula = LakeData_July$temperature_C ~ LakeData_July$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 ***
## LakeData_July$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 results suggest that depth explains 73.87% of the variability in temperature (per the r-squared value). The finding is based on 9726 degrees of freedom. The result is statistically significant at 99% as the p-value is less than .001. In this scenario we would reject the null hypothesis that depth has no impact on temperature in favor of the alternative hypothesis that temperature changes with depth.

   Based on the graph we observe that for a 1 meter increase in depth, temperature decreases by 2 degrees Celcius. From the y-intercept we can see that the mean surface temperature is 21 degrees C.

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

# setting up function that considers year4, daynum, and depth as explanatory variables
AIC_Temp <- lm(data = LakeData_July,temperature_C ~ year4 + daynum + depth)

# calling AIC in step-wise algorithm
step(AIC_Temp)
```

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

```
#10
# running a MLR on AIC recommended variables
MLR_Temp <- lm(data = LakeData_July, temperature_C ~ depth + year4 + daynum)

# displaying results
summary(MLR_Temp)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = LakeData_July)
##
## 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 AIC method suggests that it is best to keep all three explanatory variables in the model (depth, year, daynum) in order to best predict temperature. The model explains 74.12% of the variance in temperature, which is a slight improvement over using only depth as an explanatory variable. Using only depth as an explanatory variable resulted in an R-squared of 0.7387, meaning depth explained 73.87% of the variability in temperature.

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

# running ANOVA
ANOVA_Temp <- aov(data = LakeData_July, temperature_C ~ lakename)
summary(ANOVA_Temp)
```

```
##               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
```

```
# Formatting as linear model
ANOVA_Temp_asLM <- lm(data = LakeData_July, temperature_C ~ lakename)
summary(ANOVA_Temp_asLM)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = LakeData_July)
##
## 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: Since the p-value (2e-16) is less than 0.001, we reject the null hypothesis (that there is no significnat difference in mean temperature) in favor of the alternative that there is a significant difference in mean temperature among the lakes.

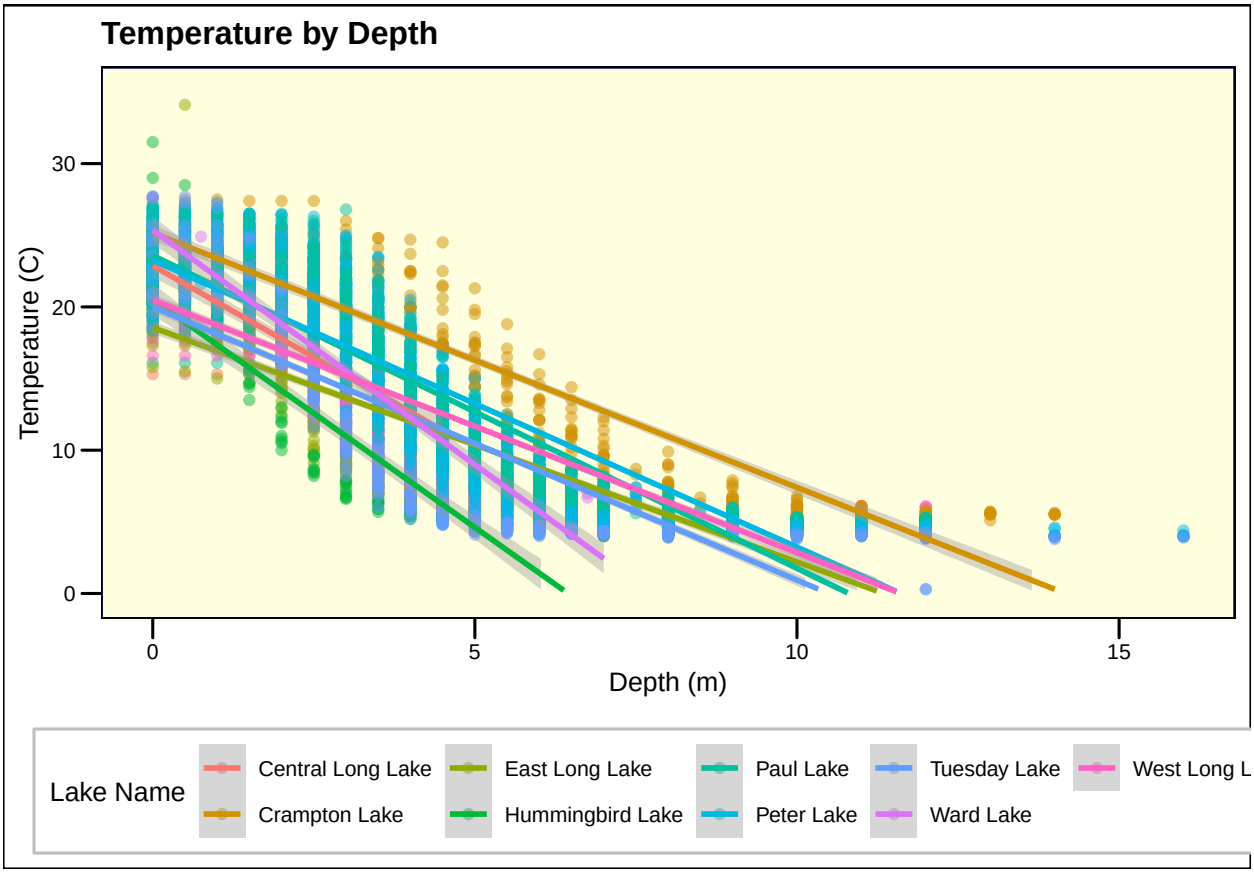
14. Create a graph that depicts temperature by depth, with a separate color for each lake. Add a geom_smooth (method = "lm", se = FALSE) for each lake. Make your points 50 % transparent. Adjust your y axis limits to go from 0 to 35 degrees. Clean up your graph to make it pretty.

```
#14.
# graphing temperature by depth for each lake
Lake_Plot_Q14 <- ggplot(LakeData_July,aes(
  x=depth,y=temperature_C,color=lakename
))+
  geom_point(alpha=0.5)+ # adding transparency
  geom_smooth(method=lm, SE=FALSE)+ # adding geom smooth
  ylim(0,35)+ # adjusting axis
  # adding plot title and axis labels
  labs(x="Depth (m)",
       y="Temperature (C)",
       title="Temperature by Depth",
       color="Lake Name")
```

```
## Warning in geom_smooth(method = lm, SE = FALSE): Ignoring unknown parameters:
## `SE`
```

```
# displaying results
Lake_Plot_Q14
```

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

## Warning: Removed 73 rows containing missing values (`geom_smooth()`).
```



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

```
#15
# using Tukey HSD test with ANOVA
LakeTemp_Tukey_Test <- HSD.test(ANOVA_Temp,"lakename",group=T)

# displaying results
LakeTemp_Tukey_Test
```

```
## $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: Statistically speaking, Paul Lake and Ward Lake have the same mean temperature as Peter Lake (all there are belong to group c). None of the lakes have a mean temperature that is statistically distinct. Based on the groupings below we can see that all lake mean temperatures overlap in some way.

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-sample "t-test" to test if there is a difference in mean temperatures between the two lakes. In this scenario, the null hypothesis would be "mean temperatures of Peter Lake and Paul Lake are equal" and the alternative hypothesis would be "mean temperatures of Peter and Paul are distinct."

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
# filtering data for just Crampton and Ward lakes
LakeData_July_Crampton_Ward <- filter(LakeData_July,lakename=="Crampton Lake"|
        lakename=="Ward Lake")

# stating null and alternative hypotheses
## H0: mu == 0
## Ha: mu != 0

# running 2-sample T-test
Crampton_Ward_2WTTEST <- t.test(
  LakeData_July_Crampton_Ward$temperature_C ~
    LakeData_July_Crampton_Ward$lakename)

# displaying result
Crampton_Ward_2WTTEST
```

```
##
##  Welch Two Sample t-test
##
## data:  LakeData_July_Crampton_Ward$temperature_C by LakeData_July_Crampton_Ward$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 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: According to the test, there is not a statistically significant difference in mean temperatures between Crampton Lake and Ward Lake. Because the p-value is not less than 0.05, we would fail to reject the null hypothesis as there is not strong evidence that a difference in means exists. This answer aligns with the output in Q16, which showed that the mean temperatures of Crampton and Ward Lakes belonged to the same statistical group of means (both overlap in group b). While the mean temp of Crampton Lake (~15 degrees C) does not technically equal the mean temp of Ward Lake (~14 degrees C), statistically speaking there is not a difference.