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

Xuancheng Guo

Spring 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.

```
# Load packages
getwd()
```

```
## [1] "/home/guest/EDA_Spring2024"
```

```
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(dplyr)
library(lubridate)
library(ggplot2)
library(here)
```

```
## here() starts at /home/guest/EDA_Spring2024
```

```r
# Load data sets
NTL.LTER <- read.csv(here("Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv"),
                     stringsAsFactors = TRUE)
NTL.LTER$sampledate <- as.Date(NTL.LTER$sampledate, format = "%m/%d/%y")
class(NTL.LTER$sampledate)
```

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

2. Build a ggplot theme and set it as your default theme.

```r
# Build Theme
mytheme <- theme_classic(base_size = 14) +
  theme(
    axis.text = element_text(color = "black"),
    legend.position = "top",
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    plot.margin = margin(10, 10, 10, 10)
  )

# Call out mytheme
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 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)

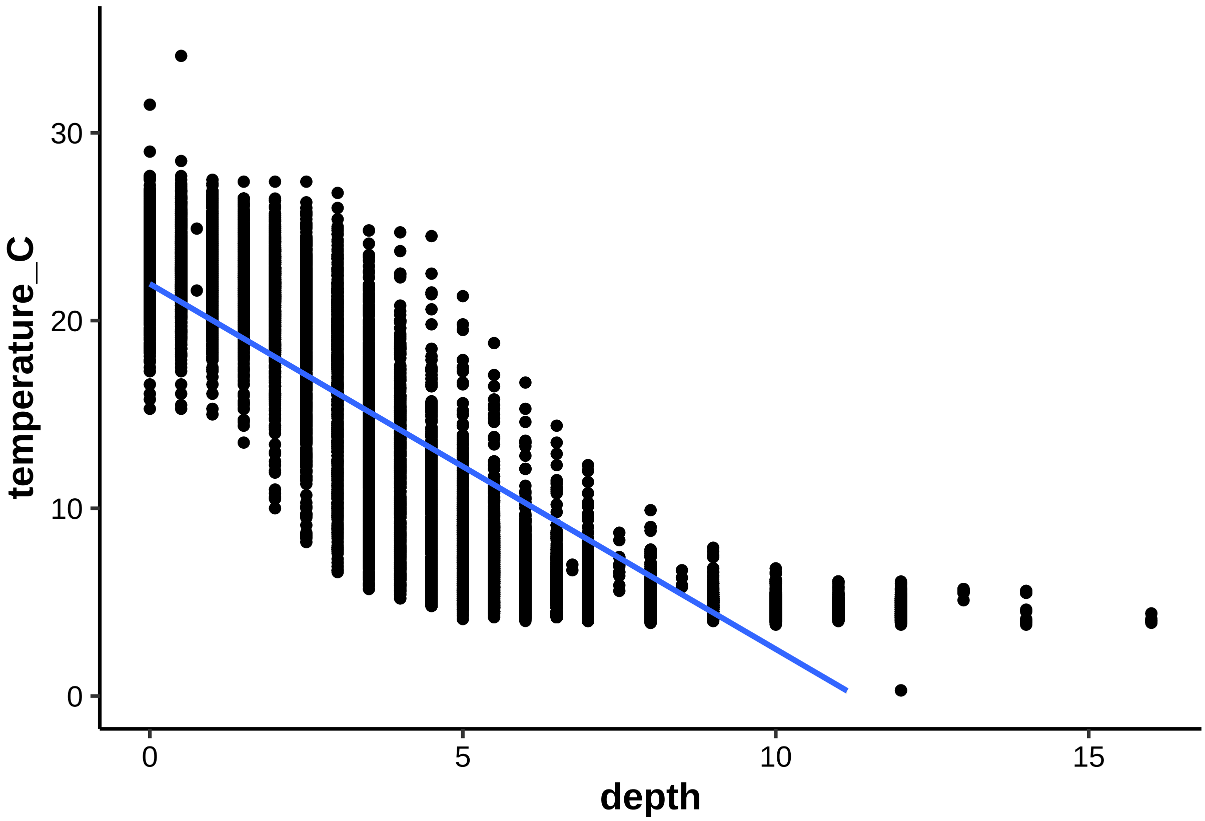
5. Visualize the relationship among the two continuous variables with a scatter plot of temperature by depth. Add a smoothed line showing the linear model, and limit temperature values from 0 to 35 °C. Make this plot look pretty and easy to read.

```r
#4
NTL.LTER.new <- NTL.LTER %>%
  mutate(month = month(sampledate)) %>%
  filter(month == 7) %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  na.omit()

#5
temp.dep <- ggplot(NTL.LTER.new, aes(x = depth, y = temperature_C)) +
  geom_point () +
  geom_smooth(method = lm) +
  ylim(0, 35)
print(temp.dep)
```

```
## ‘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: The figure suggests that as depth increase, temperature of corresponding area will decrease, but the linear prediction migth not be very accurate since the result shows a curved drop; also, the linear prediciton cannot capture result exceed 11 (roughly).

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

```
#7
temp.dep.reg <- lm(
  data = NTL.LTER.new,
  temperature_C ~ depth
  )
summary(temp.dep.reg)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = NTL.LTER.new)
##
## 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: Looking at R-square/adjusted R-square value, 73.87% of the observation is explained by the linear model presented previously. The degrees of freedom is n-2 which is summarized by the code, 9726. The result is statistically significant in such we will reject the null hypothesis at a P-value approaching 0 and conclude "the mean lake temperature recorded during July changes with depth across all lakes." By looking at point estimates of depth, we have a statistically significant result, as depth increase by 1 unit below surface, temperature will be decrease by 1.946 units.

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
temp.dep.AIC <- lm(
  data = NTL.LTER.new,
  temperature_C ~ year4 + daynum + depth
)

step(temp.dep.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 = NTL.LTER.new)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##    -8.57556      0.01134      0.03978     -1.94644
```

```
#10
summary(temp.dep.AIC)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL.LTER.new)
##
## 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: We should use all three input as variable for this regression since all three coefficien has a statistically significant level at 1% confidence level. This result is an improvement than only use depth as the explanatory variable.

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
temp.anova <- aov(data = NTL.LTER,
  temperature_C ~ lakename
  )
summary(temp.anova)
```

```
##                Df  Sum Sq Mean Sq F value Pr(>F)    
## lakename        8   57921    7240   155.7 <2e-16 ***
## Residuals   34747 1615571      46                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 3858 observations deleted due to missingness
```

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

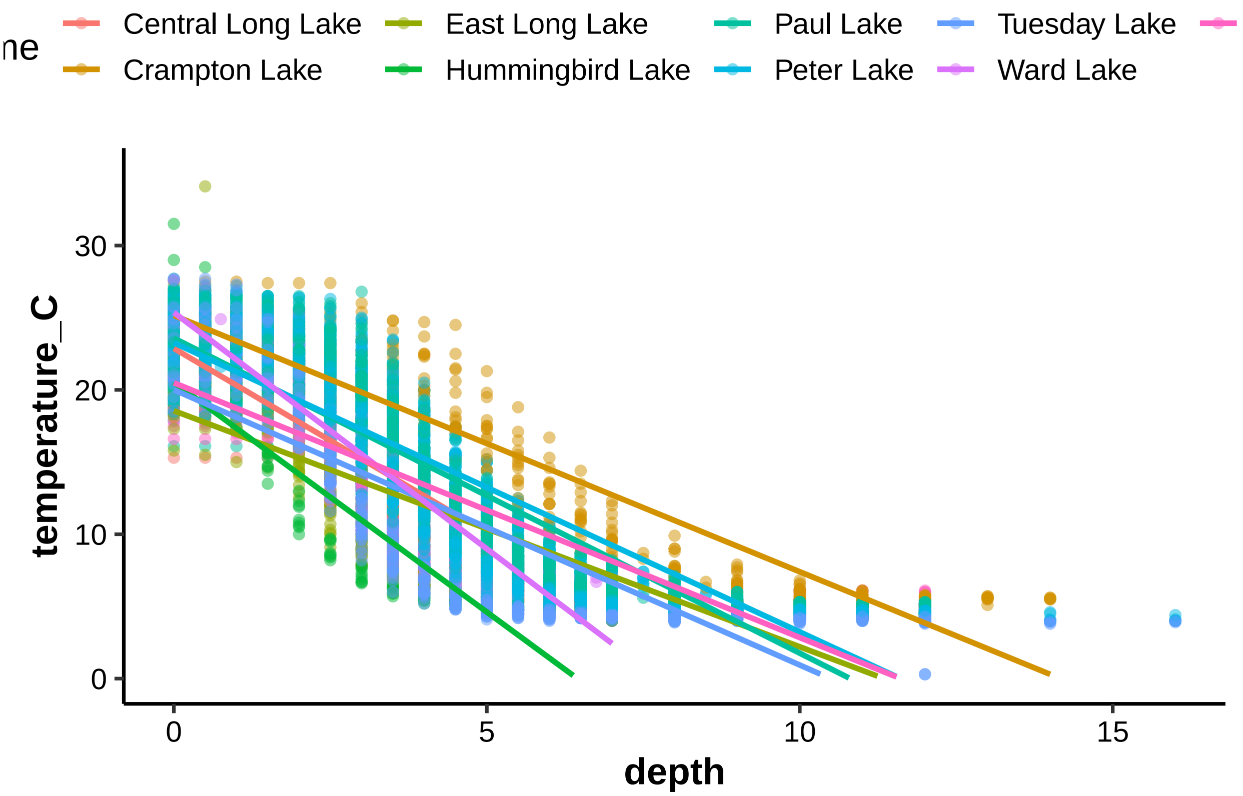
Answer: There is significant difference in temperature among lakes due to the small test P-value, so we reject the null hypothesis and conclude there is at least 1 difference among lakes.

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.
q14.graph <- ggplot(NTL.LTER.new,
  aes(x = depth, y = temperature_C, color = lakename)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  ylim(0, 35) +
  mytheme
print(q14.graph)
```

```
## `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
temp.HSD <- HSD.test(temp.anova, "lakename", group = T)
print(temp.HSD)
```

```
## $statistics
##    MSerror    Df     Mean       CV
##   46.49528 34747 11.80871 57.74336
##
## $parameters
##    test   name.t ntr StudentizedRange alpha
##   Tukey lakename   9         4.386509  0.05
##
## $means
##                    temperature_C      std     r         se Min  Max  Q25  Q50
## Central Long Lake      16.736343 4.540842   443 0.32396833 1.3 27.9 13.3 16.7
## Crampton Lake          14.192058 6.801706  1108 0.20484933 4.8 27.5  7.1 13.8
## East Long Lake          9.779296 6.304109  3550 0.11444327 3.8 34.1  4.9  6.4
## Hummingbird Lake       10.037831 6.117160   378 0.35071838 4.0 31.5  5.1  6.9
## Paul Lake              12.792275 6.783047  9253 0.07088644 3.9 27.7  6.0 11.3
## Peter Lake             12.252557 7.119817 10189 0.06755207 0.7 27.2  5.2 10.2
## Tuesday Lake           10.346702 7.027998  5503 0.09191887 0.3 27.7  4.4  6.4
## Ward Lake              12.428083 6.575945   527 0.29702918 5.0 27.6  6.6  9.9
## West Long Lake         11.058581 6.555168  3805 0.11054194 4.0 27.9  5.4  7.7
##                      Q75
## Central Long Lake 20.35
## Crampton Lake     20.80
## East Long Lake    14.70
## Hummingbird Lake  14.70
## Paul Lake         19.50
## Peter Lake        19.40
## Tuesday Lake      17.00
## Ward Lake         18.20
## West Long Lake    17.40
##
## $comparison
## NULL
##
## $groups
##                    temperature_C groups
## Central Long Lake      16.736343      a
## Crampton Lake          14.192058      b
## Paul Lake              12.792275      c
## Ward Lake              12.428083     cd
## Peter Lake             12.252557      d
## West Long Lake         11.058581      e
## Tuesday Lake           10.346702      f
## Hummingbird Lake       10.037831     fg
## East Long Lake          9.779296      g
##
## attr(,"class")
## [1] "group"
```

```
TukeyHSD(temp.anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = NTL.LTER)
##
## $lakename
##                                          diff        lwr        upr     p adj
## Crampton Lake-Central Long Lake    -2.5442854 -3.7331780 -1.3553927 0.0000000
## East Long Lake-Central Long Lake   -6.9570473 -8.0227648 -5.8913298 0.0000000
## Hummingbird Lake-Central Long Lake -6.6985124 -8.1794348 -5.2175900 0.0000000
## Paul Lake-Central Long Lake        -3.9440682 -4.9727040 -2.9154324 0.0000000
## Peter Lake-Central Long Lake       -4.4837864 -5.5102613 -3.4573116 0.0000000
## Tuesday Lake-Central Long Lake     -6.3896413 -7.4341675 -5.3451152 0.0000000
## Ward Lake-Central Long Lake        -4.3082596 -5.6715463 -2.9449730 0.0000000
## West Long Lake-Central Long Lake   -5.6777623 -6.7395105 -4.6160141 0.0000000
## East Long Lake-Crampton Lake       -4.4127620 -5.1405823 -3.6849417 0.0000000
## Hummingbird Lake-Crampton Lake     -4.1542271 -5.4140285 -2.8944256 0.0000000
## Paul Lake-Crampton Lake            -1.3997828 -2.0721371 -0.7274286 0.0000000
## Peter Lake-Crampton Lake           -1.9395011 -2.6085446 -1.2704575 0.0000000
## Tuesday Lake-Crampton Lake         -3.8453560 -4.5417779 -3.1489341 0.0000000
## Ward Lake-Crampton Lake            -1.7639743 -2.8831342 -0.6448143 0.0000357
## West Long Lake-Crampton Lake       -3.1334769 -3.8554727 -2.4114812 0.0000000
## Hummingbird Lake-East Long Lake     0.2585349 -0.8857499  1.4028198 0.9987916
## Paul Lake-East Long Lake          3.0129792  2.5954288  3.4305296 0.0000000
## Peter Lake-East Long Lake         2.4732609  2.0610627  2.8854591 0.0000000
## Tuesday Lake-East Long Lake       0.5674060  0.1121132  1.0226989 0.0035472
## Ward Lake-East Long Lake          2.6487877  1.6614645  3.6361109 0.0000000
## West Long Lake-East Long Lake     1.2792850  0.7857610  1.7728091 0.0000000
## Paul Lake-Hummingbird Lake        2.7544443  1.6446129  3.8642756 0.0000000
## Peter Lake-Hummingbird Lake       2.2147260  1.1068972  3.3225548 0.0000000
## Tuesday Lake-Hummingbird Lake     0.3088711 -0.8157039  1.4334461 0.9952041
## Ward Lake-Hummingbird Lake        2.3902528  0.9647057  3.8157999 0.0000071
## West Long Lake-Hummingbird Lake   1.0207501 -0.1198389  2.1613391 0.1224797
## Peter Lake-Paul Lake             -0.5397183 -0.8434372 -0.2359993 0.0000013
## Tuesday Lake-Paul Lake           -2.4455731 -2.8056140 -2.0855323 0.0000000
## Ward Lake-Paul Lake              -0.3641914 -1.3113688  0.5829859 0.9582889
## West Long Lake-Paul Lake         -1.7336941 -2.1410071 -1.3263812 0.0000000
## Tuesday Lake-Peter Lake          -1.9058549 -2.2596746 -1.5520351 0.0000000
## Ward Lake-Peter Lake              0.1755268 -0.7693033  1.1203570 0.9997136
## West Long Lake-Peter Lake        -1.1939759 -1.5958003 -0.7921515 0.0000000
## Ward Lake-Tuesday Lake            2.0813817  1.1169709  3.0457925 0.0000000
## West Long Lake-Tuesday Lake       0.7118790  0.2659563  1.1578017 0.0000259
## West Long Lake-Ward Lake         -1.3695027 -2.3525401 -0.3864652 0.0005266
```

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: Paul Lake and Ward Lake is likely to have the same temperature; Tuesday Lake and Hummingbird Lake also don't have a significant difference; Hummingbird Lake is also not statistically different from East Long Lake, but East Long Lake is different from Tuesday Lake. We could tell this simply from the report or look at all cross-comparing P-values, whichever P-value is large, means we cannot reject the null hypothesis that they have difference. Except the three pairs mentioned above, all other lakes are distinct than each other.

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 can run two-sample T-test or ANOVA to test the mean difference.

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
# Filter out our data
Cramp.Ward <- NTL.LTER.new %>%
  filter(lakename == "Crampton Lake" | lakename == "Ward Lake")
summary(Cramp.Ward$lakename)
```

```
## Central Long Lake     Crampton Lake    East Long Lake  Hummingbird Lake
##                 0               318                 0                 0
##         Paul Lake        Peter Lake      Tuesday Lake         Ward Lake
##                 0                 0                 0               116
##    West Long Lake
##                 0
```

```r
# Two sample T
ttest <- t.test(Cramp.Ward$temperature_C ~ Cramp.Ward$lakename)
print(ttest)
```

```
##
##  Welch Two Sample t-test
##
## data:  Cramp.Ward$temperature_C by Cramp.Ward$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: Because the P-value is large than 95% confidence level, we fail to reject that the two lakes have significant difference. This result is contradictory from Q16. One guess could be in Q16, we cross-compared all lakes, so this method first define what is “difference”, but Q18 we only performed a two sample T which we will compare our result with simple statistics.