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

Kamil Burak Karayel

Spring 2025

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(agricolae)
library(here)
```

```
## here() starts at /home/guest/EDA_Spring2025_kbk
```

```r
library(lubridate)

getwd()
```

```
## [1] "/home/guest/EDA_Spring2025_kbk"
```

```r
here()
```

```
## [1] "/home/guest/EDA_Spring2025_kbk"
```

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

LakeData$sampledate <- mdy(LakeData$sampledate)
class(LakeData$sampledate)
```

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

```r
#2
library(ggthemes)

theme_kbk <- theme_base() +
  theme(
    plot.background = element_rect(colour = 'black', fill = 'grey'),
    #background is grey and frame is black
    plot.title = element_text(size = 15, colour = 'red'),
    #title of the plot is red and size of 15
    axis.title = element_text(size = 12, face = "bold", colour = "darkred"),
    #axis labels are dark red, bold and size of 12
    axis.text = element_text(size = 10, face = "italic"),
    #indicators of the axis are italic and size of 10
    legend.position = 'bottom'
    #legend will be at the bottom of the plot
  )

theme_set(theme_kbk)   #set my theme as default
```

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: There is no significant difference in mean lake temperature recorded during July between different depths across all lakes. Ha: There is significant difference in mean lake temperature recorded during July between different depths 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.

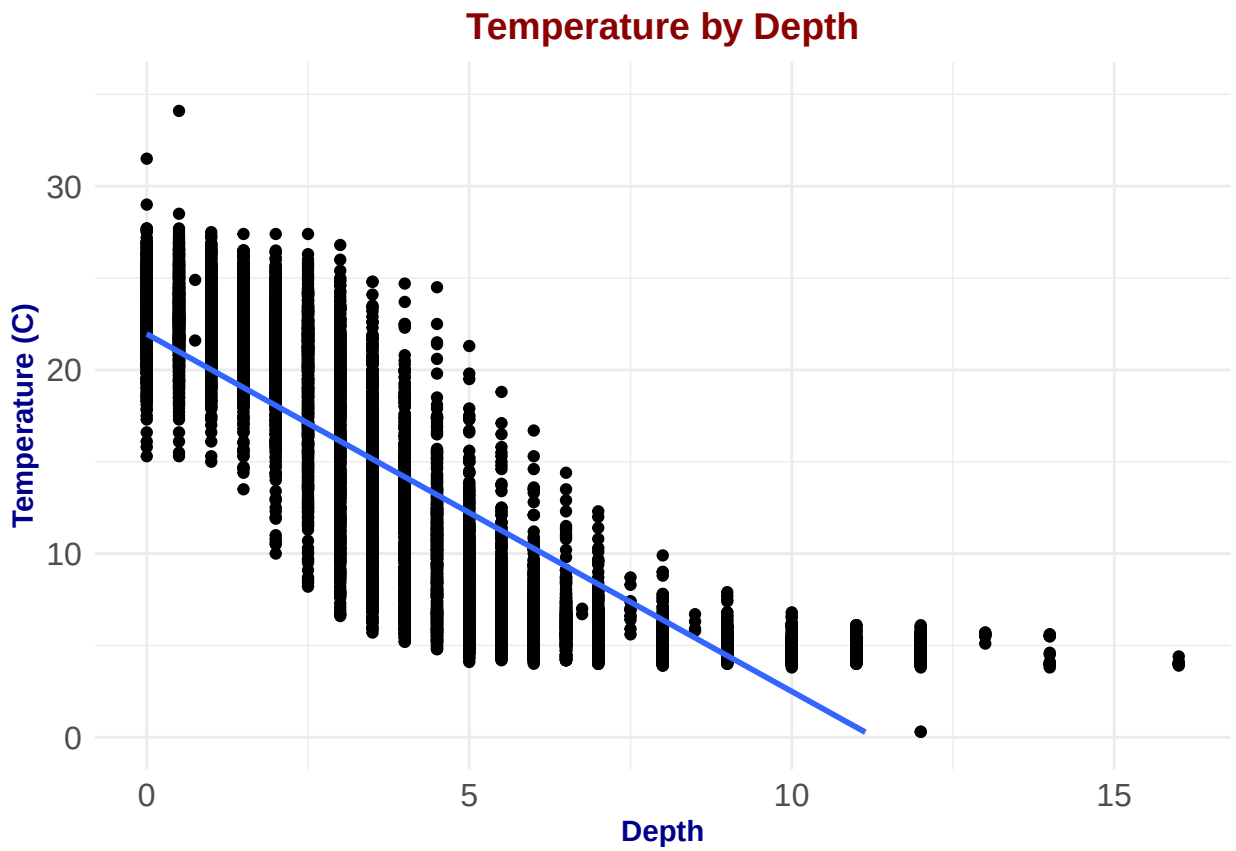
```
#4
Q4 <- LakeData %>%
  filter(format(sampledate, "%m")=="07") %>%  #filter data in July where "m" of sampledate is "07"
  select(lakename:daynum, depth, temperature_C) %>%  #select lakename to daynum, depth and temperature
  drop_na()    #drop incomplete cases

#5
Q5 <- ggplot(data = Q4, aes(x=depth, y=temperature_C))+
  geom_point() +     #scatterplot of temperature vs depth
  geom_smooth(method = "lm") +     #add smoothed line
  ylim(0,35) +   #limit temperatures
  labs(title = "Temperature by Depth",   #add axis titles
       x="Depth",
       y="Temperature (C)") +
  theme_minimal() +
  theme(       #add some visual developments
    axis.title = element_text(face = "bold", colour = "darkblue"),
    axis.text = element_text(size = 12),
    plot.title = element_text(size = 14, hjust = 0.5, face = "bold", colour = "darkred")
  )

Q5   #show the plot
```

```
## `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: We can say by the graph that the differences between actual points and linear model is high, especially depths under 10. There may not be a purely linear relationship between depth and temperature.

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

```
#7
Q7 <- lm(data = Q4, temperature_C ~ depth)   #linear regression of temperature by depth

summary(Q7)   #show the information about regression
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = Q4)
##
## 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: With 73.87% value of R-squared, we can say that changes in depth play a significant role in temperatures. P value is smaller than 0.05, even it is nearly zero. This means that relationship between depth and temperature is statistically significant. And the coefficient for depth is -1.95, this means 1 meter depth causes 1.95 degrees decrease in temperature.

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
Q9_AIC <- lm(data = Q4, temperature_C ~ year4 + daynum + depth)   #run an AIC to determine necessary
                                                                  #and adequate variables.
step(Q9_AIC)   #AIC value is 26066. By excluding year4 AIC is 26070,
```

```
## 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 = Q4)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##    -8.57556      0.01134      0.03978     -1.94644
```

```
                #by excluding daynum AIC is 26148, by excluding depth AIC is 39189.
                #Excluding any variable results higher AIC. So we need to include all three variables

#10
Q10 <- lm(data = Q4, temperature_C ~ year4 + daynum + depth)    #include all variables
                                                                #to the multiple regression

summary(Q10)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = Q4)
##
## 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: AIC value of including all three variables is 26066. By excluding year4 AIC is 26070, by excluding daynum AIC is 26148, by excluding depth AIC is 39189. Excluding any variable results higher AIC. So AIC method suggests to use year4, daynum and depth in our multiple regression. R squared value was 73.87% for only depth. But now, R-squared value increases to 74.12%. This means year4 and daynum contribute to prediction of 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
Q12_anova <- aov(data = Q4, temperature_C ~ lakename) #run anova to compare mean temperatures among lak

summary(Q12_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
```

```
Q12_linear <- lm(data = Q4, temperature_C ~ lakename)

summary(Q12_linear)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = Q4)
##
## 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: Findings from ANOVA says that at least one lake has sinificantly different mean temperatures than the other lakes, because the p-value is close to zero(<2e-16). Linear model confirms that result with near zero p-values for every lakes. That means every lakes mean temperatures are significantly different.

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.

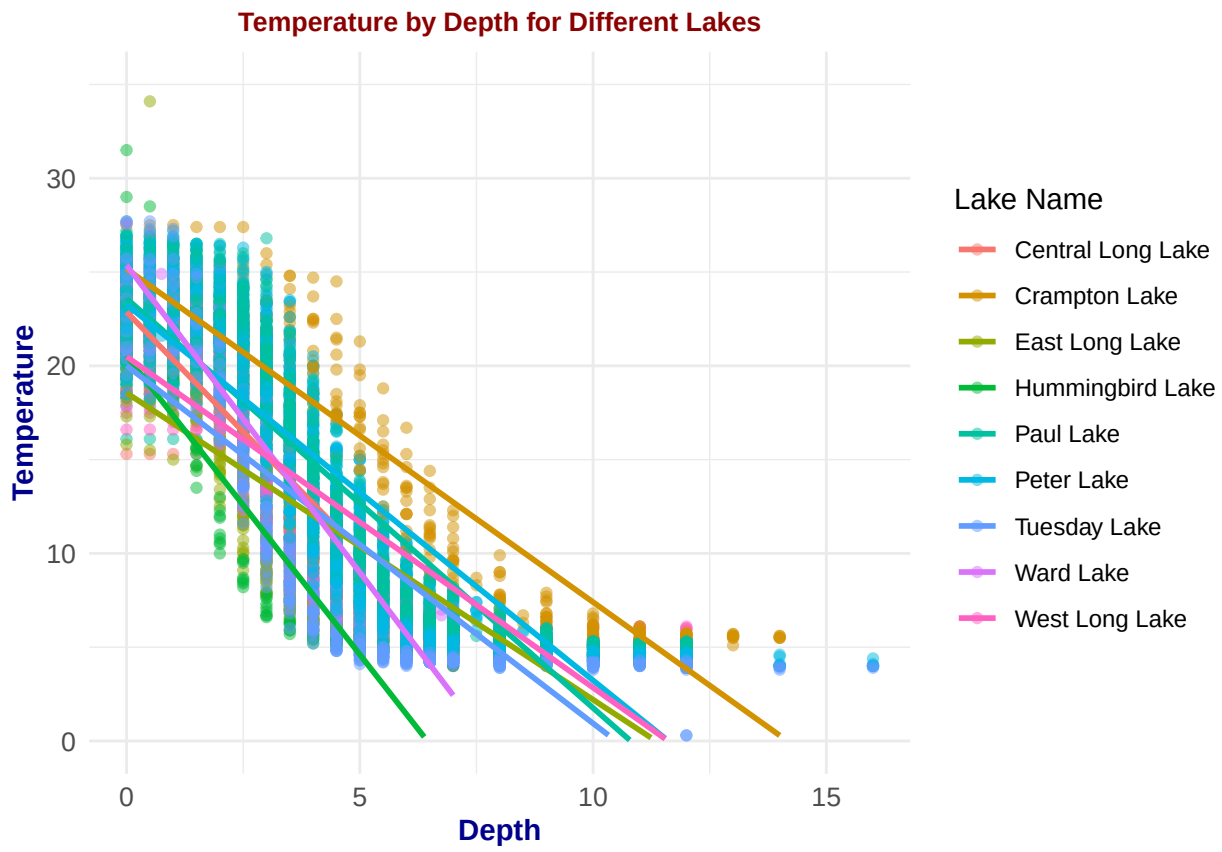
```
#14.
Q14 <- ggplot(Q4, aes(x=depth, y=temperature_C, color=lakename)) +
  geom_point(alpha=0.5) +    #50% transparent points
  geom_smooth(method = "lm", se=FALSE) +
  ylim(0,35) +   #display between 0 and 35 degrees
```

```
  labs(
    x="Depth",
    y="Temperature",
    title = "Temperature by Depth for Different Lakes",
    color="Lake Name"
  ) +
  theme_minimal() +
  theme(       #add some visual developments
    axis.title = element_text(face = "bold", colour = "darkblue"),
    axis.text = element_text(size = 10),
    plot.title = element_text(size = 10, hjust = 0.5, face = "bold", colour = "darkred"))

Q14
```

```
## `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.

```
#15
Q15_tukey <- HSD.test(Q12_anova, "lakename", group = TRUE) #Tukey's HSD test for ANOVA from Q12

Q15_tukey
```

```
## $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: Peter Lake falls into group "c" in the Tukey's HSD results. Statistically Paul Lake (c) and Ward Lake (bc) don't have significantly different mean temperatures from Peter Lake.

Central Long Lake is in group "a" and does not share a letter with any other lake except Crampton Lake (ab), meaning it is significantly warmer than most lakes but not completely distinct. Similarly East Long Lake is in group "e" and does not share any letters with other lakes except Tuesday Lake and Hummingbird Lake (de), meaning it is significantly colder than most lakes but not entirely distinct. Since no lake has a completely unique group letter, no lake is entirely distinct from all the others, though Central Long Lake and East Long Lake are the most extreme in temperature.

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: In order to determine whether Peter and Paul Lakes have distinct mean temperatures, we might use independent t-test.

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?

```
Q18_crampton <- Q4%>%    #make a new dataset from Q4 which includes only temperature values of Crampton
  filter(lakename == "Crampton Lake") %>%  #take only data from Crampton Lake
  select(temperature_C) %>%  #take only temperature data
  drop_na()

Q18_ward <- Q4%>%    #make a new dataset from Q4 which includes only temperature values of Ward Lake
  filter(lakename == "Ward Lake") %>%   #take only data from Ward Lake
  select(temperature_C) %>%   #take only temperature data
  drop_na()

Q18_ttest <- t.test(Q18_crampton, Q18_ward)    #run two sample t test form these datasets

Q18_ttest
```

```
##
##  Welch Two Sample t-test
##
## data:  Q18_crampton and Q18_ward
## t = 1.1181, df = 200.37, p-value = 0.2649
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.6821129  2.4686451
## sample estimates:
## mean of x mean of y
##  15.35189  14.45862
```

Answer: p-value is greater than 0.05 and confidence interval includes 0. By these values of t-test, we can reject the null hypothesis where the mean temperatures of Crampton and Ward Lakes are equal. Hence there is not significant difference between mean temperatures of these Lakes in July. In Q16 Crampton Lake was in group (ab) and Ward Lake was in group (bc). That means these are not equal. This difference may be due to different assumptions of dşfferent tests. Tukey's HSD compares various lakes at the same time but t-test only compares two lakes.