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

Andi Mujollari

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
setwd("/Users/andimujollari/Desktop/EDE-Fall2023/")
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
```

```
#install.packages("agricolae")
library(agricolae)
raw_data <- read.csv("./Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv")
```

```
raw_data$sampledate <- as.Date(raw_data$sampledate, format = "%m/%d/%Y")

getwd()
```

```
## [1] "/Users/andimujollari/Desktop/EDE-Fall2023"
```

```
#2
# Here i set the theme

library(ggplot2)

custom_theme <- theme_minimal() +
  theme(
    text = element_text(size = 14, color = "black"),
    axis.title = element_text(face = "bold"),
    axis.text = element_text(color = "green"),
    panel.background = element_rect(fill = "lightgray"),
    panel.grid.major = element_line(color = "gray"),
    panel.grid.minor = element_blank()
  )
theme_set(custom_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: The mean lake temperature recorded during July doesn't 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.

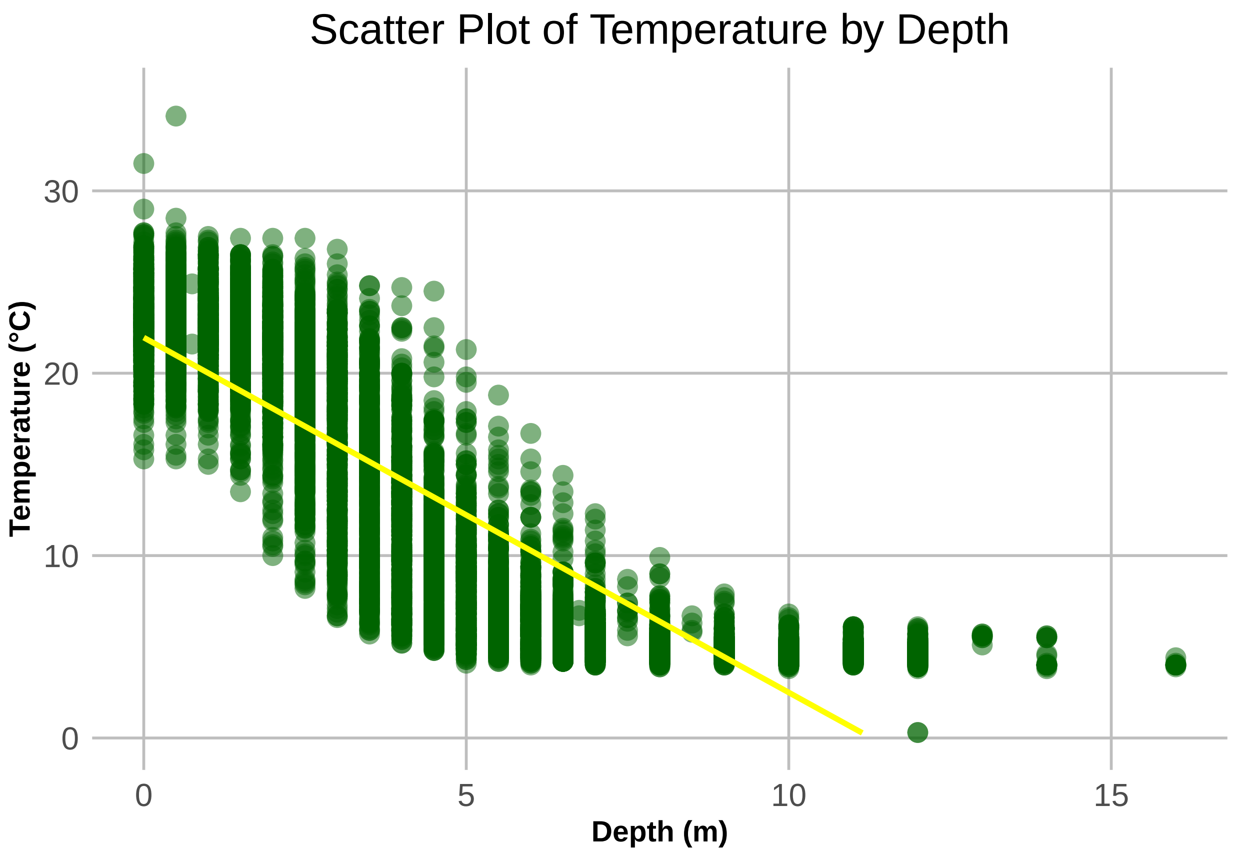
```
#4
library(dplyr)
library(tidyr)

# Here I filter the dates to show only July, select specific columns, and remove NAs
filtered_data <- raw_data %>%
  filter(format(sampledate, "%m") == "07") %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  drop_na()

# View the first few rows of the filtered dataset
head(filtered_data)
```

```
##    lakename year4 daynum depth temperature_C
## 1 Paul Lake  1984    183   0.0          22.8
## 2 Paul Lake  1984    183   0.5          22.9
## 3 Paul Lake  1984    183   1.0          22.8
## 4 Paul Lake  1984    183   1.5          22.7
## 5 Paul Lake  1984    183   2.0          21.7
## 6 Paul Lake  1984    183   2.5          20.3
```

```r
#5

library(ggplot2)

scatter_plot <- ggplot(data = filtered_data, aes(x = depth, y = temperature_C)) +
  geom_point(color = "darkgreen", size = 3, alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, color = "yellow") +

  # Customize my plot's appearance
  labs(
    title = "Scatter Plot of Temperature by Depth",
    x = "Depth (m)",
    y = "Temperature (°C)"
  ) +
  theme_minimal() +
  theme(
    panel.grid.major = element_line(color = "gray"),
    panel.grid.minor = element_blank(),
    axis.title = element_text(face = "bold"),
    axis.text = element_text(size = 12),
    plot.title = element_text(size = 16, hjust = 0.5)
  ) +
  ylim(0, 35)

# Print the scatter plot
print(scatter_plot)
```

```
## ‘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 is a negative correlation between temperature and depth across every lake. The higher the depth, the lower the temperature.

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

```
#7
# Linear regression model
linear_model <- lm(temperature_C ~ depth, data = filtered_data)

# Here i display the summary of the linear regression model
summary(linear_model)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = filtered_data)
##
## 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: From the results of the linear regression we notice that almost 73.87% of the variability in temperature is explained by the change in depth.This results are extracted by analizing our sample of 9726 datapoints (degrees of freedom). Which means the results of the model are statistically significant as we can see from the p-values that are almost zero. From the results we notice that for every meter increase of the depth, the temperature will decrease by 1.94621 grace Celcius.

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


# Here I create a list of candidate models with different combinations of predictor variables
candidate_models <- list(
  model1 = lm(temperature_C ~ year4, data = filtered_data),
  model2 = lm(temperature_C ~ daynum, data = filtered_data),
  model3 = lm(temperature_C ~ depth, data = filtered_data),
  model4 = lm(temperature_C ~ year4 + daynum, data = filtered_data),
  model5 = lm(temperature_C ~ year4 + depth, data = filtered_data),
  model6 = lm(temperature_C ~ daynum + depth, data = filtered_data),
  model7 = lm(temperature_C ~ year4 + daynum + depth, data = filtered_data)
)

#Using
model7 = lm(temperature_C ~ year4 + daynum + depth, data = filtered_data)
step(model7)
```

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

```
# Calculate AIC for each model
AIC_values <- sapply(candidate_models, AIC)

# Find the model with the lowest AIC
best_model <- names(AIC_values)[which.min(AIC_values)]

# Display AIC values and the best model
AIC_values
```

```
##   model1   model2   model3   model4   model5   model6   model7
## 66819.14 66796.54 53762.12 66798.34 53756.97 53679.36 53674.39
```

```
best_model
```

```
## [1] "model7"
```

```
#10

# Herei will try to create a multiple regression model with year4, daynum, and depth as predictors
multiple_regression_model <- lm(temperature_C ~ year4 + daynum + depth, data = filtered_data)

# And display the summary of the multiple regression model
summary(multiple_regression_model)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = filtered_data)
##
## 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 that the AIC method suggest we use to predict temperature are year4, daynum and depth. these explain 74.12% of the observe temperature variance. When compared to the model that used only depth the model with three variables has an improvment by 0.25% of the observed temperature variance.

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

# First I run an ANOVA test to compare mean temperatures among lakes
anova_model <- aov(temperature_C ~ lakename, data = filtered_data)

# Display the summary of the ANOVA model
summary(anova_model)
```

```
##               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
```

```
# Here i it a linear model with lakename as a predictor
linear_model_lakes <- lm(temperature_C ~ lakename, data = filtered_data)

# Display the summary of the linear model
summary(linear_model_lakes)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = filtered_data)
```

```
##
## 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.

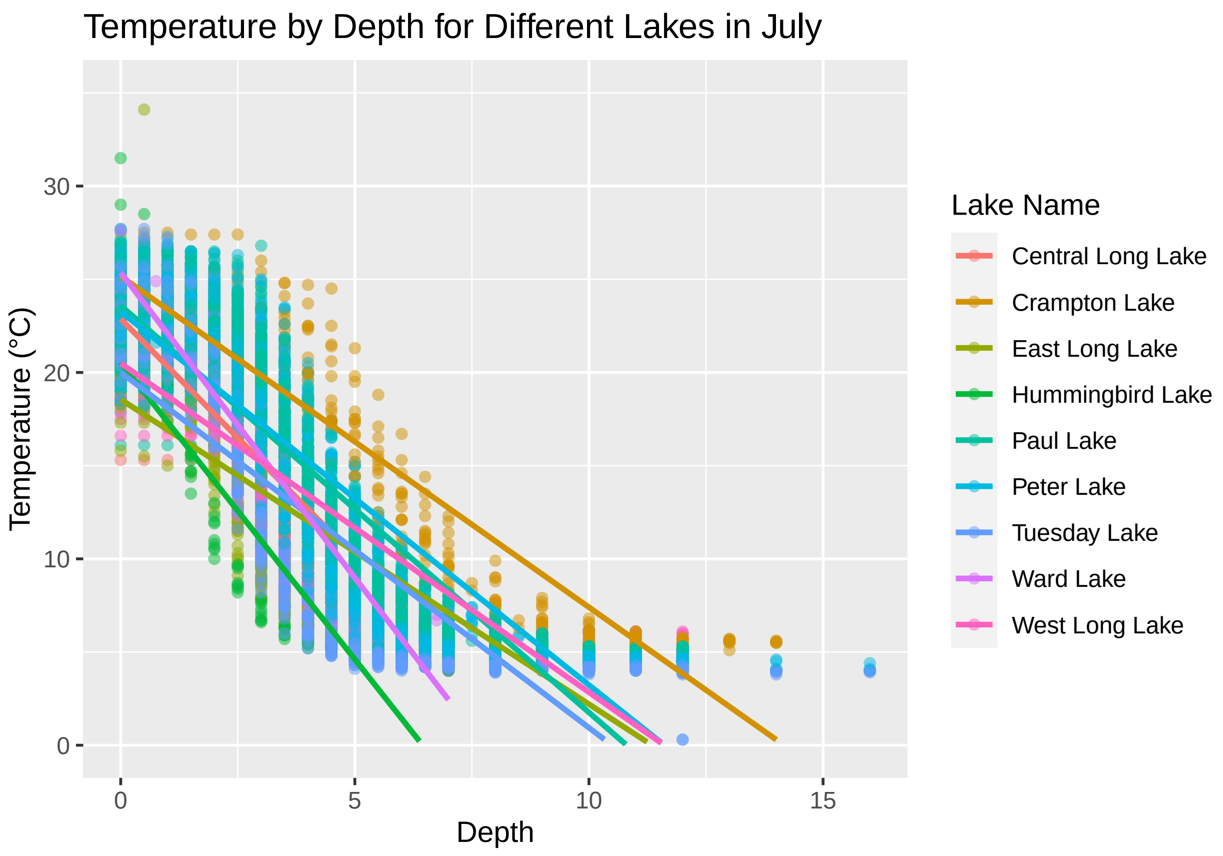
    Answer: Yes, we observe a significant difference in the mean temperature for the month of July. Which is explained by both the ANOVA test and linear model. P-values of th elinear model are almost zero, which indicates the difference in mean temperature of each of the lakes are statistically significant.

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.

```r
#14.

# Here i create the scatter plot with separate colors for each lake
ggplot(filtered_data, aes(x = depth, y = temperature_C, color = lakename)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  ylim(0, 35) +
  labs(
    x = "Depth",
    y = "Temperature (°C)",
    title = "Temperature by Depth for Different Lakes in July",
    color = "Lake Name"
  ) +
  theme_gray()
```

```
## `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

# Load stats package
library(stats)

# here i perform Tukey's HSD test to compare means of different lakes
tukey_result <- HSD.test(aov(temperature_C ~ lakename, data = filtered_data),"lakename", group = T)

# Finally i print the results
print(tukey_result)
```

```
## $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: The means of Peter Lake and Crampton Lake, Hummingbird Lake, Tuesday Lake, and West Long Lake are statistically distinct, as indicated by the adjusted p-values being less than 0.05. However, there is no statistically significant difference in means between Peter Lake and Paul Lake or between Peter Lake and Ward Lake, as the adjusted p-values are greater than 0.05.

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: Another test that will help us to explore if the mean temperatures are distinct is the independent samples 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?

```r
# Subset the data for Crampton Lake and Ward Lake
crampton_lake_data <- filtered_data[filtered_data$lakename == "Crampton Lake", "temperature_C"]
ward_lake_data <- filtered_data[filtered_data$lakename == "Ward Lake", "temperature_C"]

# Perform a two-sample t-test
t_test_result <- t.test(crampton_lake_data, ward_lake_data)

# Print the t-test result
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  crampton_lake_data and ward_lake_data
## 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: The Welch Two-Sample t-test results show that there isn't enough evidence to suggest that the mean temperatures for Crampton Lake and Ward Lake in July are significantly different. However, this contradicts the outcome of Tukey's HSD test in question 16, which showed significant differences in mean temperatures between lakes. The difference could be due to the disparity in the statistical tests used. The t-test focuses on a specific comparison between two lakes, while Tukey's HSD test considers multiple pairwise comparisons among lakes. These differences highlight the importance of selecting appropriate statistical tests and considering the specific research question and data characteristics.