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

Angelica Rodriguez

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

#Setting the directory

setwd('/home/guest/EDA_Spring2025/Data/Raw')

#Intalling packages and calling libraries

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 1.9.4     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
```

```r
library(here)
```

```
## here() starts at /home/guest/EDA_Spring2025
```

```r
library(lubridate)
library(cowplot)
```

```
##
## Attaching package: 'cowplot'
##
## The following object is masked from 'package:lubridate':
##
##     stamp
```

```r
#install.packages("agricolae")
library(agricolae)
library(ggplot2)

#confirming the library
getwd()
```

```
## [1] "/home/guest/EDA_Spring2025/Data/Raw"
```

```r
#to import the database

raw_data <- read.csv("NTL-LTER_Lake_ChemistryPhysics_Raw.csv")

#establishing column sampledate as a date
str(raw_data)
```

```
## 'data.frame':    38614 obs. of  11 variables:
##  $ lakeid         : chr  "L" "L" "L" "L" ...
##  $ lakename       : chr  "Paul Lake" "Paul Lake" "Paul Lake" "Paul Lake" ...
##  $ year4          : int  1984 1984 1984 1984 1984 1984 1984 1984 1984 1984 ...
##  $ daynum         : int  148 148 148 148 148 148 148 148 148 148 ...
##  $ sampledate     : chr  "5/27/84" "5/27/84" "5/27/84" "5/27/84" ...
##  $ depth          : num  0 0.25 0.5 0.75 1 1.5 2 3 4 5 ...
##  $ temperature_C  : num  14.5 NA NA NA 14.5 NA 14.2 11 7 6.1 ...
##  $ dissolvedOxygen: num  9.5 NA NA NA 8.8 NA 8.6 11.5 11.9 2.5 ...
##  $ irradianceWater: num  1750 1550 1150 975 870 610 420 220 100 34 ...
##  $ irradianceDeck : num  1620 1620 1620 1620 1620 1620 1620 1620 1620 1620 ...
##  $ comments       : chr  NA NA NA NA ...
```

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

#checking if sample date is a date object

class(raw_data$sampledate)
```

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

```r
#2

Angelica_custom_theme_2 <- function() {
theme_minimal() +
        theme(
      plot.background = element_rect(fill = "#f8f9fa", color = NA),
            panel.background = element_rect(fill = "#ffffff", color = NA),
      plot.title = element_text(color = "#005f73", size = 18, face = "bold", hjust = 0.5),
      axis.title = element_text(color = "#005f73", size = 14, face = "italic"),
      axis.text = element_text(color = "#4a4a4a", size = 12),
            axis.line = element_line(color = "#468faf", linewidth = 1),
      axis.ticks = element_line(color = "#4a4a4a", linewidth = 1),
      legend.background = element_rect(fill = "#ffffff", color = NA),
      legend.title = element_text(color = "darkblue", size = 12, face = "bold"),
      legend.text = element_text(color = "darkred", size = 10),
      panel.grid.major = element_line(color = "gray90", linetype = "dashed"),
      panel.grid.minor = element_blank()
    )
}



# I set my custom theme as the default theme
theme_set(Angelica_custom_theme_2())
```

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 in all lakes. - The mean temperature is independent of depth.- H1: The mean lake temperature recorded during July changes with depth in at least one lake.-The mean temperature varies as a function of depth-

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

# Load the dplyr and tidyr packages
library(dplyr)
library(tidyr)

# Here I filter for records with dates in July, select specific columns, and remove NAs
```

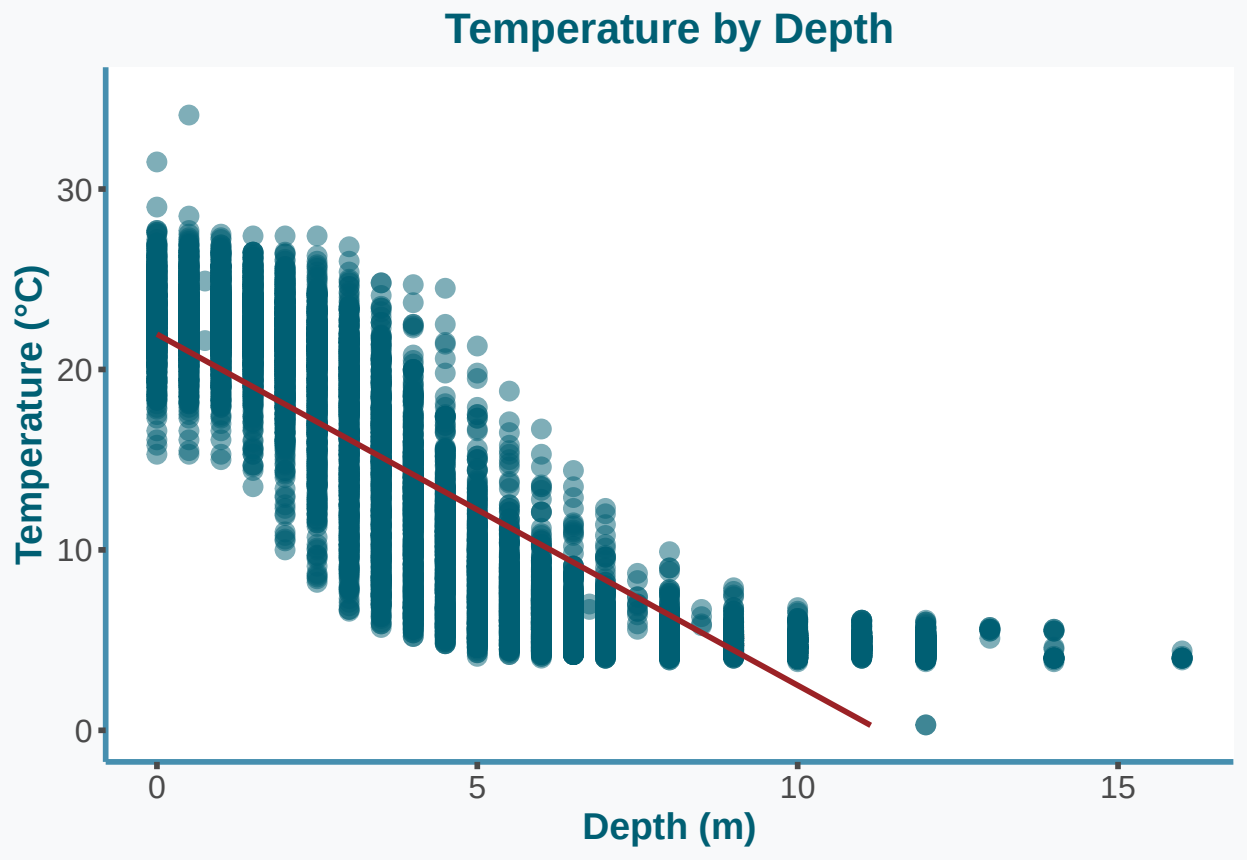
```r
NTL_LTER_clean <- raw_data %>%
filter(format(sampledate, "%m") == "07") %>%
select(lakename, year4, daynum, depth, temperature_C) %>%
drop_na()

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

```
##    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
library(dplyr)

#5

#Creating the Scatter plot: Temperature and Depth
Scatter_plot_Temp_Depth <- ggplot(NTL_LTER_clean, aes(x = depth, y = temperature_C)) +
geom_point(color = "#005f73", size = 3, alpha = 0.5) +
geom_smooth(method = "lm", color = "#9b2226", se = FALSE) +

# Labels of the graphic
labs(
title = "Temperature by Depth",
x = "Depth (m)",
y = "Temperature (°C)") +
#Design of the plot
Angelica_custom_theme_2()+
theme(
panel.grid.major = element_blank(),
panel.grid.minor = element_blank(),
axis.title = element_text(face = "bold"),
axis.text = element_text(size = 12),
plot.title = element_text(face = "bold",size = 16, hjust = 0.5)
) +
ylim(0, 35)

print(Scatter_plot_Temp_Depth)
```

```
## `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: Actually the the figrue shows that there is not much linearity in the distribution of the data, by the contrary it looks like it has a curve distribution.

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

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

#to see the results

summary(linear_model)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = NTL_LTER_clean)
##
## 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:

The intercept is 21.95597, meaning that if the depth equals 0, the water temperature is 21.96 °C. In the case of the slope (-1.94621), for each additional meter of depth, the water temperature decreases by 1.95 °C on average. These data show that there is a compensation between temperature and depth. In the case of the range of errors, the minimum is -9.52, the maximum is 13.58, and the median is 0.067, which means that it is very close to 0, indicating that this model does not have a very high bias. On the other hand, the 1st Quartile (-3.02 °C) and 3rd Quartile (2.94 °C) indicate that the central 50% of the errors are in this range, which gives an idea of the dispersion that, despite there being variability in the errors, this variability is not significant. Finally, both coefficients are highly significant (***), meaning a strong relationship exists between temperature and 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.

```r
#9

 #to create different models
candidate_models <- list(
model_full <- lm(temperature_C ~ year4 + daynum + depth, data = NTL_LTER_clean),
# Model without year4
model_no_year <- lm(temperature_C ~ daynum + depth, data = NTL_LTER_clean),
# Model without daynum
model_no_day <- lm(temperature_C ~ year4 + depth, data = NTL_LTER_clean),
# Model without depth
model_no_depth <- lm(temperature_C ~ year4 + daynum, data = NTL_LTER_clean),
```

```r
# Model with only depth (since it seems important)
model_only_depth <- lm(temperature_C ~ depth, data = NTL_LTER_clean),
# Null model (intercept only, no predictors)
model_null <- lm(temperature_C ~ 1, data = NTL_LTER_clean))

AIC_values <- data.frame(
  Model = c("Full Model", "No Year", "No Day", "No Depth", "Only Depth", "Null Model"),
  AIC = c(AIC(model_full), AIC(model_no_year), AIC(model_no_day),
          AIC(model_no_depth), AIC(model_only_depth), AIC(model_null)))

# Sort models by AIC value (lowest AIC = best model)
AIC_values <- AIC_values[order(AIC_values$AIC), ]

# Print the ordered AIC values
print(AIC_values)
```

```
##        Model      AIC
## 1 Full Model 53674.39
## 2    No Year 53679.36
## 3     No Day 53756.97
## 5 Only Depth 53762.12
## 4   No Depth 66798.34
## 6 Null Model 66817.36
```

```r
# Identify the best model (lowest AIC)
best_model_name <- AIC_values$Model[1]  # Model with the lowest AIC
best_model <- candidate_models[[best_model_name]]
cat("The best model based on AIC is:", best_model_name, "\n")
```

```
## The best model based on AIC is: Full Model
```

```r
#Best model is:model_full


#10

#running the best model

multiple_regression_model <- lm(temperature_C ~ year4 + daynum + depth, data = NTL_LTER_clean)

summary(multiple_regression_model)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL_LTER_clean)
##
## 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: A) temperature_C = year4 + daynum + depth is the final set of explanatory variables that the AIC method suggests. B) According to the multiple regression table, the model explains 74.12% of the observed variance in water temperature. This is reflected in the R^2 value (Multiple R-squared = 0.7412), which indicates what proportion of the variability in temperature can be predicted by the variables year4, daynum, and depth. C) For this analysis, it is best to compare the R^2 of this multiple regression with the model that only uses depth. Assuming that the model with only depth had an R^2 of 0.7387, then: -The model with only depth $R^2 = 0.7387$ -The model with year4, daynum, and depth $R^2 = 0.7412$ This difference in R^2 demonstrates an improvement, though it is not very significant. The R^2 increased from 0.7387 to 0.7412, indicating that adding year4 and daynum contributes slightly to explaining the temperature better, but not significantly.

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).

```r
#12

#ANOVA model

anova_model <- aov(temperature_C ~ lakename, data = NTL_LTER_clean)

# 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
```

```r
#linear model

linear_model <- lm(temperature_C ~ lakename, data = NTL_LTER_clean)

# Display the summary of the linear model

summary(linear_model)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = NTL_LTER_clean)
##
## 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: According to the analysis of variances, ANOVA confirms that there is a significant difference between the temperature of the lakes (50 <2e-16 ***). This variance $p < 0.05$ shows that it is possible to reject the null hypothesis, which states that all lakes have the same average temperature. However, the linear regression shows that the temperature of the lakes is not significantly different among the lakes so it is necesary to include other types of variables that can show a more explanations to the data.

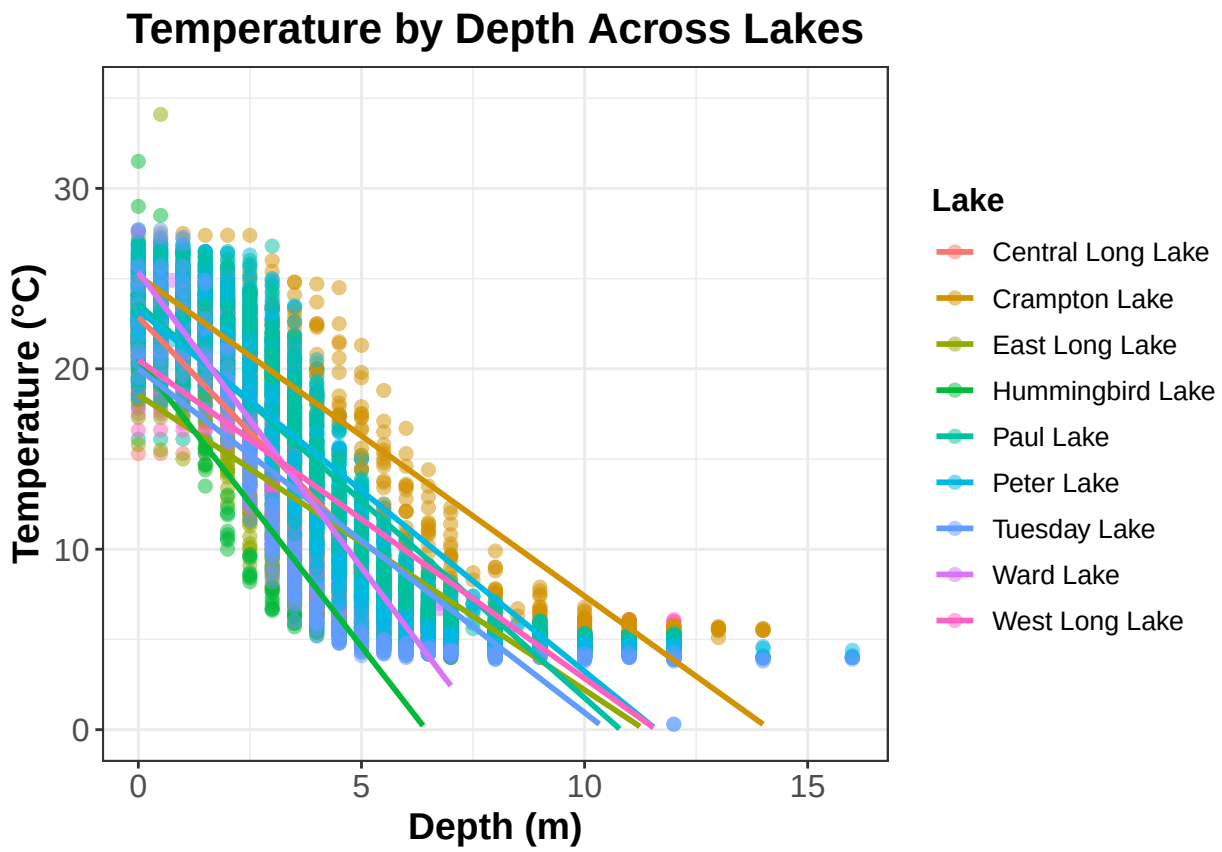
14. Create a graph that depicts temperature by depth, with a separate color for each lake. Add a geom_smooth (method = "lm", se = FALSE) for each lake. Make your points 50 % transparent. Adjust your y axis limits to go from 0 to 35 degrees. Clean up your graph to make it pretty.

```r
#14.

#Creating the plot: Scatter points with 50% transparency

ggplot(NTL_LTER_clean, aes(x = depth, y = temperature_C, color = lakename)) +
```

```
  geom_point(alpha = 0.5, size = 2) +

# Linear regression for each lake framewok of the plot
  geom_smooth(method = "lm", se = FALSE, linewidth = 1) +
  labs(
    title = "Temperature by Depth Across Lakes",
    x = "Depth (m)",
    y = "Temperature (°C)",
    color = "Lake"
  ) +
  theme_bw() +  # Clean theme
  theme(
    plot.title = element_text(size = 16, face = "bold", hjust = 0.5),
    axis.title = element_text(size = 14, face = "bold"),
    axis.text = element_text(size = 12),
    legend.title = element_text(size = 12, face = "bold"),
    legend.text = element_text(size = 10)
  ) +
  ylim(0, 35)
```

```
## `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
library(stats)
# Perform Tukey's HSD test to compare means of different lakes
tukey_test_result <- HSD.test(aov(temperature_C ~ lakename, data = NTL_LTER_clean),"lakename", group = 
# Print the results
print(tukey_test_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 lakes have the same mean temperature as Peter Lake is Paul Lake and with a bit diference can be Ward Lake. According to the Tukey test, the lakes with the most significant difference in mean temperatures are Central Long Lake, with a mean temperature of 17.66, and East Long Lake, with 10.26. Thus, it can be observed that Central Long Lake is warmer than East Long Lake.

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: T test. Tukey test is made for multiple comparissons but if we want to se the difference between just two lakes T test Is the best option to evaluate if the mean tempweratures are statistically significant.

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?

```
crampton_lake_data <- NTL_LTER_clean [NTL_LTER_clean$lakename == "Crampton Lake", "temperature_C"]

ward_lake_data <- NTL_LTER_clean [NTL_LTER_clean$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: For this analysis, the following hypotheses I worked where: H0: There is no difference in the average temperatures between Crampton Lake and Ward Lake. H1: The average temperatures of the lakes are different. According to the T-test, having a T score of 1.1181 compared to the P value of 0.2699, there is insufficient evidence to reject the null hypothesis. From this it can be concluded that there is no significant difference between the temperature of Crampton Lake and Ward Lake. Are the average temperatures of the lakes the same? On the other hand,

the average of both lakes is different since the T-test shows that the average of Crampton Lake is 15.35189, and the average of Ward Lake is 14.45862. However, the difference between both means it is very small, which makes the P value not show such a significant difference. Finally, this analysis compared bye the analysis in point 16, the answer varies slightly. Since, according to the Tukey test analysis, Paul Lake has an average temperature close to Ward Lake and Peter Lake, Crampton Lake was not considered in this analysis. On the other hand, the T-test and Tukey test both agree that there is no significant difference in temperature between Crampton Lake and Ward Lake, and both lakes in the two tests have the same mean.

Table 1: Crampton Lake y Ward Lake

Lake	temperature_C	groups	T-test mean
Crampton Lake	15.35189	ab	15.35189
Ward Lake	14.45862	bc	14.45862