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

Chrissie Pantoja

Fall 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 necessary packages
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   1.5.0
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.1
## -- 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 /Users/chrissiepantoja/Library/CloudStorage/OneDrive-DukeUniversity/PHD DUKE/1 COURS
```

```r
#Import EPA data (from the processed_KEY folder) & fix dates
ntl_lter_data <- read.csv(
  here("Data", "Raw", "NTL-LTER_Lake_ChemistryPhysics_Raw.csv"),
  stringsAsFactors = TRUE)

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

# Check the structure of the data to identify date columns
str(ntl_lter_data)
```

```
## 'data.frame':    38614 obs. of  11 variables:
##  $ lakeid         : Factor w/ 9 levels "C","E","H","L",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ lakename       : Factor w/ 9 levels "Central Long Lake",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ 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     : Date, format: "1984-05-27" "1984-05-27" ...
##  $ 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       : Factor w/ 2 levels "DO Probe bad - Doesn't go to zero",..: NA NA NA NA NA NA NA 
```

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

```r
# Load necessary library
library(ggplot2)

# Create a custom ggplot theme
custom_theme <- theme_minimal() +
  theme(
    text = element_text(family = "Arial", size = 12, color = "black"),
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5, size = 14),
    axis.title.x = element_text(size = 14),
    axis.title.y = element_text(size = 14),
    axis.text = element_text(size = 12),
    legend.position = "bottom",
    legend.title = element_text(size = 12),
    legend.text = element_text(size = 10),
    panel.grid.major = element_line(color = "grey80"),
    panel.grid.minor = element_line(color = "grey90"),
    panel.border = element_blank(),
    plot.background = element_rect(fill = "white", color = NA)
  )

# Set the custom theme as the default
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: - Null Hypothesis (Ho): There is no relationship between mean lake temperature in July and depth across all lakes. - Alternative Hypothesis (Ha): There is a relationship between mean lake temperature in July and 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)

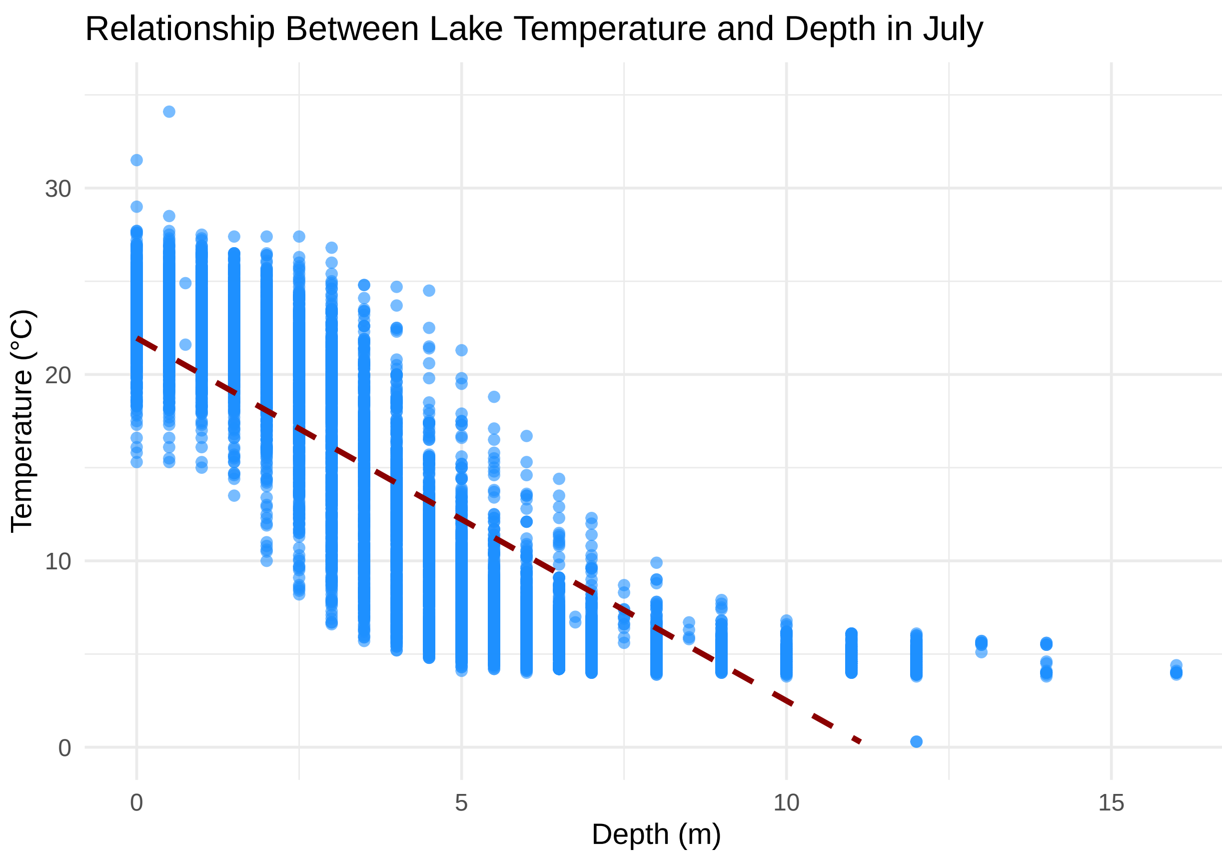
```
ntl_july_data <- ntl_lter_data %>%
  filter(month(sampledate) == 7) %>%  # Keep only July records
  select(lakename, year4, daynum, depth, temperature_C) %>%
  na.omit()  # Remove rows with NA values
```

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.

```
ggplot(ntl_july_data, aes(x = depth, y = temperature_C)) +
  geom_point(alpha = 0.6, color = "dodgerblue") +  # Scatter plot with points
  geom_smooth(method = "lm", se = FALSE, color = "darkred", linetype = "dashed") +
  labs(
    title = "Relationship Between Lake Temperature and Depth in July",
    x = "Depth (m)",
    y = "Temperature (°C)"
  ) +
  scale_y_continuous(limits = c(0, 35)) +
  theme_minimal()
```

```
## `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: The figure shows a clear negative relationship between lake temperature and depth in July. As depth increases, the temperature generally decreases. This is a common trend in lakes due to the way sunlight penetrates and warms the water.The distribution of points suggests that the relationship is mostly linear. The points cluster around a downward-sloping line, indicating that the decrease in temperature with depth is relatively consistent. However, there is also some scatter in the data, which suggests that other factors besides depth might also influence temperature.

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

```
# Perform linear regression of temperature on depth
temp_depth_lm <- lm(temperature_C ~ depth, data = ntl_july_data)

# Display the results
summary(temp_depth_lm)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = ntl_july_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: The linear regression results indicate that depth has a significant negative effect on lake temperature in July. Specifically, the coefficient for depth is -1.94621, meaning that for every 1-meter increase in depth, lake temperature is predicted to decrease by approximately 1.94°C. The model explains about 74% of the variability in temperature (R-squared = 0.7387), highlighting depth as a key factor influencing lake temperature. The residual standard error is 3.835, with 9726 degrees of freedom, and the F-statistic of 2.75 with a p-value of less than 2.2e-16 (or 0.05) confirms that this relationship is highly statistically significant

---

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.

```r
# Run the multiple regression model
TPAIC <- lm(data = ntl_july_data, temperature_C ~ year4 + daynum + depth)
summary(TPAIC)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = ntl_july_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
```

```
#Choose a model by AIC in a Stepwise Algorithm
step(TPAIC)
```

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

10. Run a multiple regression on the recommended set of variables.

```
# Run the multiple regression on the selected predictors
TPmodel <- lm(data = ntl_july_data, temperature_C ~ daynum + depth)

# Display the results of the multiple regression
summary(TPmodel)
```

```
##
## Call:
## lm(formula = temperature_C ~ daynum + depth, data = ntl_july_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6174 -2.9809  0.0845  2.9681 13.4406
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 14.088588   0.855505   16.468   <2e-16 ***
## daynum       0.039836   0.004318    9.225   <2e-16 ***
```

```
## depth       -1.946111   0.011685 -166.541   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.818 on 9725 degrees of freedom
## Multiple R-squared:  0.741,  Adjusted R-squared:  0.741
## F-statistic: 1.391e+04 on 2 and 9725 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?

```r
# Run the multiple regression on the selected predictors
test_model <- lm(data = ntl_july_data, temperature_C ~ daynum)

# Display the results of the multiple regression
summary(test_model)
```

```
##
## Call:
## lm(formula = temperature_C ~ daynum, data = ntl_july_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.320  -7.156  -2.594   8.077  21.399
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.722359   1.675347   2.819  0.00483 **
## daynum      0.040502   0.008475   4.779 1.79e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.494 on 9726 degrees of freedom
## Multiple R-squared:  0.002343,   Adjusted R-squared:  0.00224
## F-statistic: 22.84 on 1 and 9726 DF,  p-value: 1.786e-06
```

Answer: Using AIC-based stepwise selection, the final model to predict temperature includes daynum and depth as explanatory variables, excluding year4 due to its minimal contribution. This model explains 74% of the observed variance (R-squared = 0.741), with a residual standard error of 3.818, indicating a strong fit. In comparison, a model with only depth provides a lower R-squared, while a model with daynum alone explains only 0.23% of the variance (R-squared = 0.0023) and has a residual standard error of 7.494. Thus, the daynum and depth model substantially improves explanatory power, capturing far more variance than models using either predictor alone.

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
summary(ntl_july_data$lakename)
```

```
## Central Long Lake     Crampton Lake    East Long Lake  Hummingbird Lake
##               128               318               968               116
##         Paul Lake        Peter Lake      Tuesday Lake         Ward Lake
##              2660              2872              1524               116
##    West Long Lake
##              1026
```

```r
# Format ANOVA as aov
#Temperature.Totals.anova <- aov(ntl_july_data$temperature_C ~ ntl_july_data$lakename)
Temperature.Totals.anova <- aov(data = ntl_july_data, temperature_C ~ lakename)
summary(Temperature.Totals.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
```

```r
#results: reject null hypothesis i.e. difference between a pair of group means is statistically signifi

# Format ANOVA as lm
Temperature.Totals.anova2 <- lm(data = ntl_july_data, temperature_C ~ lakename)
summary(Temperature.Totals.anova2)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = ntl_july_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.

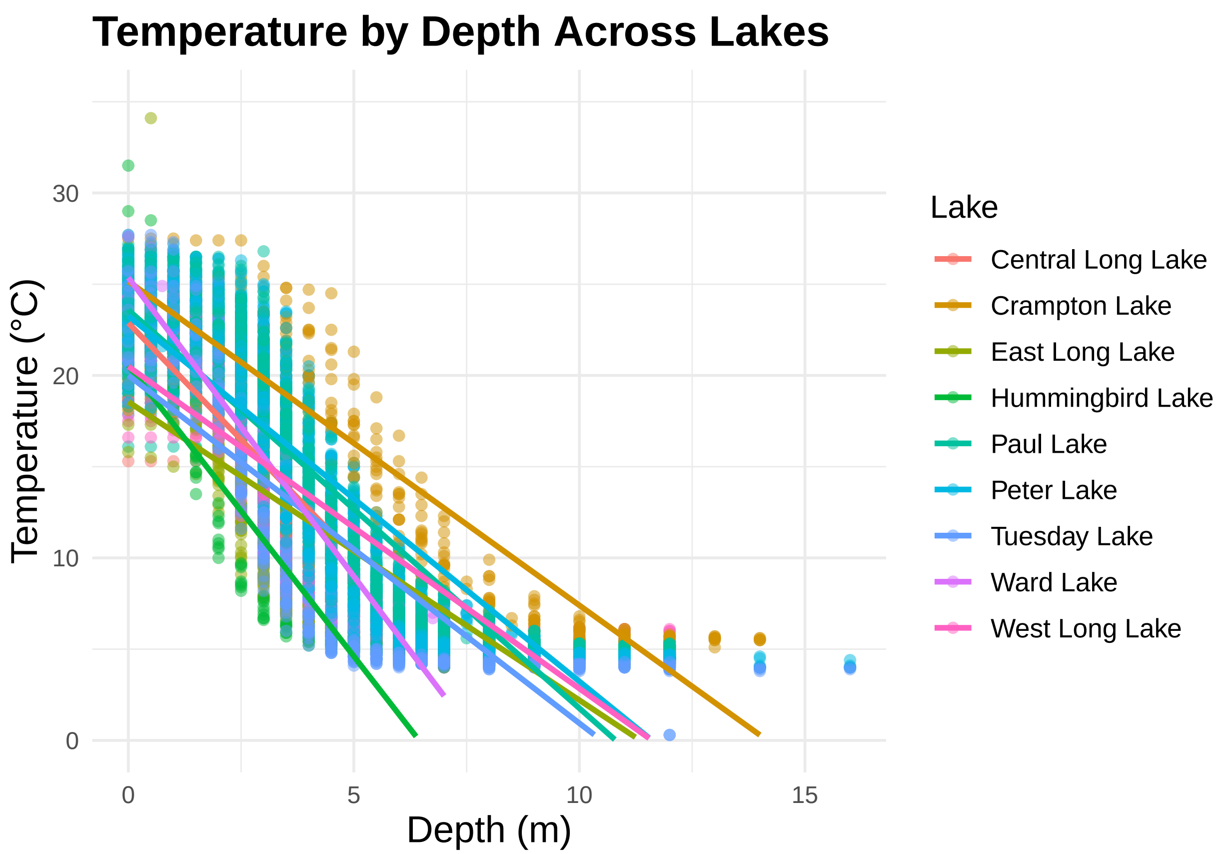
    Answer: The analysis suggests that the temperature differences among the lakes are statistically significant (p-value<0.05), as evidenced by both the ANOVA and the linear model outputs. Thus, we reject the null hypothesis that states the means of the temperatures are equal across different lakes.

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.

```
# Create the plot
ggplot(data = ntl_july_data, aes(x = depth, y = temperature_C, color = lakename)) +
  geom_point(alpha = 0.5) +  # Make points 50% transparent
  geom_smooth(method = "lm", se = FALSE) +
  scale_y_continuous(limits = c(0, 35)) +
  labs(
    title = "Temperature by Depth Across Lakes",
    x = "Depth (m)",
    y = "Temperature (°C)",
    color = "Lake"
  ) +
  theme_minimal() +  # Use a minimal theme for a clean look
  theme(
    plot.title = element_text(size = 16, face = "bold"),
    axis.title = element_text(size = 14),
    legend.title = element_text(size = 12),
    legend.text = element_text(size = 10)
  )
```

```
## `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.

```
TukeyHSD(Temperature.Totals.anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = ntl_july_data)
##
## $lakename
##                                            diff        lwr        upr     p adj
## Crampton Lake-Central Long Lake    -2.3145195 -4.7031913  0.0741524 0.0661566
## East Long Lake-Central Long Lake   -7.3987410 -9.5449411 -5.2525408 0.0000000
## Hummingbird Lake-Central Long Lake -6.8931304 -9.8184178 -3.9678430 0.0000000
## Paul Lake-Central Long Lake        -3.8521506 -5.9170942 -1.7872070 0.0000003
## Peter Lake-Central Long Lake       -4.3501458 -6.4115874 -2.2887042 0.0000000
## Tuesday Lake-Central Long Lake     -6.5971805 -8.6971605 -4.4972005 0.0000000
## Ward Lake-Central Long Lake        -3.2077856 -6.1330730 -0.2824982 0.0193405
## West Long Lake-Central Long Lake   -6.0877513 -8.2268550 -3.9486475 0.0000000
## East Long Lake-Crampton Lake       -5.0842215 -6.5591700 -3.6092730 0.0000000
## Hummingbird Lake-Crampton Lake     -4.5786109 -7.0538088 -2.1034131 0.0000004
## Paul Lake-Crampton Lake            -1.5376312 -2.8916215 -0.1836408 0.0127491
## Peter Lake-Crampton Lake           -2.0356263 -3.3842699 -0.6869828 0.0000999
## Tuesday Lake-Crampton Lake         -4.2826611 -5.6895065 -2.8758157 0.0000000
## Ward Lake-Crampton Lake            -0.8932661 -3.3684639  1.5819317 0.9714459
```

```
## West Long Lake-Crampton Lake       -3.7732318 -5.2378351 -2.3086285 0.0000000
## Hummingbird Lake-East Long Lake     0.5056106 -1.7364925  2.7477137 0.9988050
## Paul Lake-East Long Lake            3.5465903  2.6900206  4.4031601 0.0000000
## Peter Lake-East Long Lake           3.0485952  2.2005025  3.8966879 0.0000000
## Tuesday Lake-East Long Lake         0.8015604 -0.1363286  1.7394495 0.1657485
## Ward Lake-East Long Lake            4.1909554  1.9488523  6.4330585 0.0000002
## West Long Lake-East Long Lake       1.3109897  0.2885003  2.3334791 0.0022805
## Paul Lake-Hummingbird Lake          3.0409798  0.8765299  5.2054296 0.0004495
## Peter Lake-Hummingbird Lake         2.5429846  0.3818755  4.7040937 0.0080666
## Tuesday Lake-Hummingbird Lake       0.2959499 -1.9019508  2.4938505 0.9999752
## Ward Lake-Hummingbird Lake          3.6853448  0.6889874  6.6817022 0.0043297
## West Long Lake-Hummingbird Lake     0.8053791 -1.4299320  3.0406903 0.9717297
## Peter Lake-Paul Lake               -0.4979952 -1.1120620  0.1160717 0.2241586
## Tuesday Lake-Paul Lake             -2.7450299 -3.4781416 -2.0119182 0.0000000
## Ward Lake-Paul Lake                 0.6443651 -1.5200848  2.8088149 0.9916978
## West Long Lake-Paul Lake           -2.2356007 -3.0742314 -1.3969699 0.0000000
## Tuesday Lake-Peter Lake            -2.2470347 -2.9702236 -1.5238458 0.0000000
## Ward Lake-Peter Lake                1.1423602 -1.0187489  3.3034693 0.7827037
## West Long Lake-Peter Lake          -1.7376055 -2.5675759 -0.9076350 0.0000000
## Ward Lake-Tuesday Lake              3.3893950  1.1914943  5.5872956 0.0000609
## West Long Lake-Tuesday Lake         0.5094292 -0.4121051  1.4309636 0.7374387
## West Long Lake-Ward Lake           -2.8799657 -5.1152769 -0.6446546 0.0021080
```

```
GroupTukeyHSD <- HSD.test(Temperature.Totals.anova,"lakename", group = TRUE)
GroupTukeyHSD
```

```
## $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: Paul Lake and Ward Lake have statistically similar mean temperatures to Peter Lake (they falls into group "c"). No lake has a statistically distinct mean temperature from all the others.

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: To determine whether Peter Lake and Paul Lake have distinct mean temperatures, you can conduct a t-test. This statistical test compares the means of two groups and assesses whether any observed difference is 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?

```r
# Wrangle the data to include only Crampton Lake and Ward Lake
selected_lakes <- ntl_july_data %>%
  filter(lakename %in% c("Crampton Lake", "Ward Lake"))

# Conduct the two-sample t-test
t_test_result <- t.test(temperature_C ~ lakename, data = selected_lakes)

print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  temperature_C by 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: Since the p-value is greater than 0.05, the two-sample t-test between Crampton Lake and Ward Lake confirms that their mean temperatures are not statistically different in July. Therefore, this result aligns with the conclusion in part 16, supporting that no lake has a distinct mean temperature that separates it from all others statistically.