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

Victoria Thompson

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.

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

```
#1
getwd() #get working directory
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```
#load 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   1.5.1
## 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 error
```

```
library(agricolae)
library(ggplot2)
library(lubridate)
library(dplyr)


#load in raw data
NTL.raw <- read.csv("Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv",
                    stringsAsFactors = TRUE)

#put dates in date format
NTL.raw$sampledate <- mdy(NTL.raw$sampledate)

#2
mytheme <- theme_gray(base_size = 12) +
  theme(axis.text = element_text(color = "darkblue"),
        legend.position = "right",
         plot.title = element_text(face = "bold", size = 16,
                                   color = "black", hjust = 1))

# set default
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.

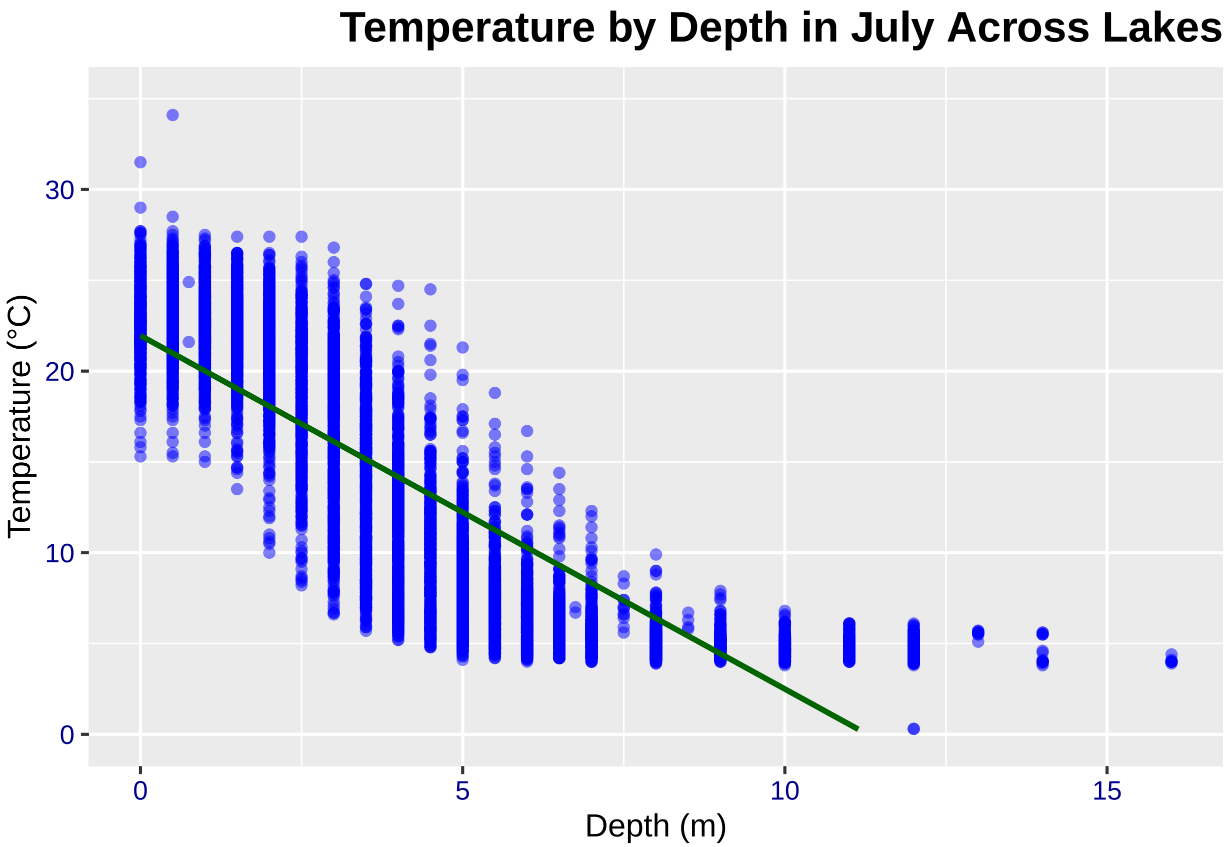
```
#4
NTL.filtered <- NTL.raw %>%
  filter(month(sampledate) == 7)  %>% #month 7 = July
  select(lakename, year4, daynum, depth, temperature_C) %>%
  # Select specific columns
  drop_na() #drop entire cases with na's
```

```
#5
NTL.temp.v.depth <- ggplot(NTL.filtered, aes(x = depth, y = temperature_C)) +
  geom_point(alpha = 0.5, color = "blue") + #transparent points to better
  #show density
  geom_smooth(method = "lm", color = "darkgreen", se = FALSE) +
  scale_y_continuous(limits = c(0, 35)) +
  labs(
    title = "Temperature by Depth in July Across Lakes",
    x = "Depth (m)",
    y = "Temperature (°C)"
  ) +
  mytheme

#print figure
NTL.temp.v.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: The figure suggests that as depth increases, temperature decreases. The points themselves are slightly curved, almost in a exponential pattern. The points are more linear at lower depths, and flatten as depth increases. This could indicate that there may be a non-linear regression that fits the data better.

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

```r
#7

#perform the regression
temp.depth.regression <-
  lm(NTL.filtered$temperature_C ~
       NTL.filtered$depth)

#display results
summary(temp.depth.regression)
```

```
##
## Call:
## lm(formula = NTL.filtered$temperature_C ~ NTL.filtered$depth)
##
## 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 ***
## NTL.filtered$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 p value of the model is <2e-16, and the Rˆ2 value is 0.7387. With such a small p-value, the likelihood that the relationship between variables is purely chance is very low. This means that the relationship is statistically significant, and that changes in depth are very likely to influence lake temperature. The Rˆ2 value of 0.7387 suggests that approximately 73.87% of the variability in lake temperature can be explained by changes in depth. This indicates a strong relationship, as depth explains a large portion of the temperature variation. There were 9726 degrees of freedom. To determine how much temperature is predicted to change for every 1m change in depth, the slope of the line of best fit model (y=mx+b; slope is “m”) is used. The coefficient of depth is -1.94621, meaning that for everying 1-meter increase in depth, the model predicts a decrease in temperature by about 1.95 degrees C.

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

library(corrplot)
```

```
## corrplot 0.94 loaded
```

```
#subset any possible explanatory variables
NTL.subset <- NTL.raw %>%
  filter(month(sampledate) == 7)  %>% #month 7 = July
  select(year4, daynum, depth, temperature_C) %>%
  na.omit()

#Run AIC
NTLAIC <- lm(data = NTL.subset, temperature_C ~ depth + year4 +
              daynum)

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

```
## Start:  AIC=26065.53
## temperature_C ~ depth + year4 + daynum
##
##          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 ~ depth + year4 + daynum, data = NTL.subset)
##
## Coefficients:
## (Intercept)        depth        year4       daynum
##    -8.57556     -1.94644      0.01134      0.03978
```

```
NTLmodel <- lm(data = NTL.subset, temperature_C ~ depth + year4 +
              daynum)
summary(NTLAIC)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = NTL.subset)
##
## 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
## depth       -1.946437   0.011683 -166.611  < 2e-16 ***
## year4        0.011345   0.004299    2.639  0.00833 **
## daynum       0.039780   0.004317    9.215  < 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
```

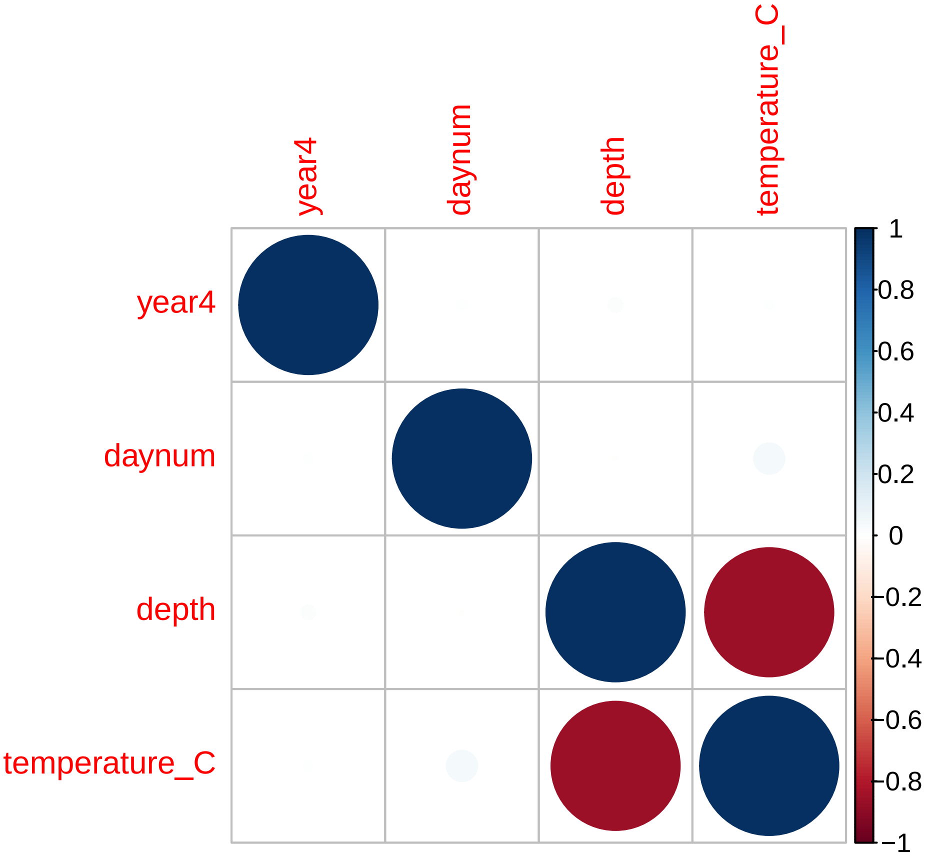
```
#create a plot to help visualize

NTLcor <- cor(NTL.subset)
corrplot(NTLcor , upper= "ellipse")
```

```
## Warning in text.default(pos.xlabel[, 1], pos.xlabel[, 2], newcolnames, srt =
## tl.srt, : "upper" is not a graphical parameter
```

```
## Warning in text.default(pos.ylabel[, 1], pos.ylabel[, 2], newrownames, col =
## tl.col, : "upper" is not a graphical parameter
```

```
## Warning in title(title, ...): "upper" is not a graphical parameter
```



```
#the plot indicates that depth and daynum may both have a role in predicting
#temperature

#10
#running a regression with both depth and irradiance
Temp.multi.regression <- lm(data = subset(NTL.raw, month(sampledate) == 7),
                   temperature_C ~ depth + daynum)
summary(Temp.multi.regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + daynum, data = subset(NTL.raw,
##     month(sampledate) == 7))
##
## 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 ***
## depth       -1.946111   0.011685 -166.541   <2e-16 ***
## daynum       0.039836   0.004318    9.225   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3.818 on 9725 degrees of freedom
##   (1116 observations deleted due to missingness)
## 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?

    Answer: The AIC suggested that the variables of daynum and depth play a role in predicting temperature. This new model explains 74.1% of variance, as derived from the new R^2 value. This means that the combination of these variables explains more of the variation than just depth alone, since the regression with just depth could only explain 73.87% of variation in 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).

```r
#12
NTL.filtered.lakename <- NTL.raw %>%
  filter(month(sampledate) == 7)  %>% #month 7 = July
  select(lakename, year4, daynum, depth, temperature_C,lakename) %>%
  # Select specific columns
  drop_na() #drop entire cases with na's

#anova
lake_anova <- aov(temperature_C ~ lakename, data = NTL.filtered.lakename)
summary(lake_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
#linear model

lake_lm <- lm(temperature_C ~ lakename, data = NTL.filtered.lakename)
summary(lake_lm)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = NTL.filtered.lakename)
##
## 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 ANOVA results indicate a statistically significant difference in mean temperature among the lakes in July, with a p-value of less than 2e-16 and an F-value of 50. This very low p-value suggests strong evidence to reject the null hypothesis, meaning that at least one lake's mean temperature is significantly different from the others.

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.

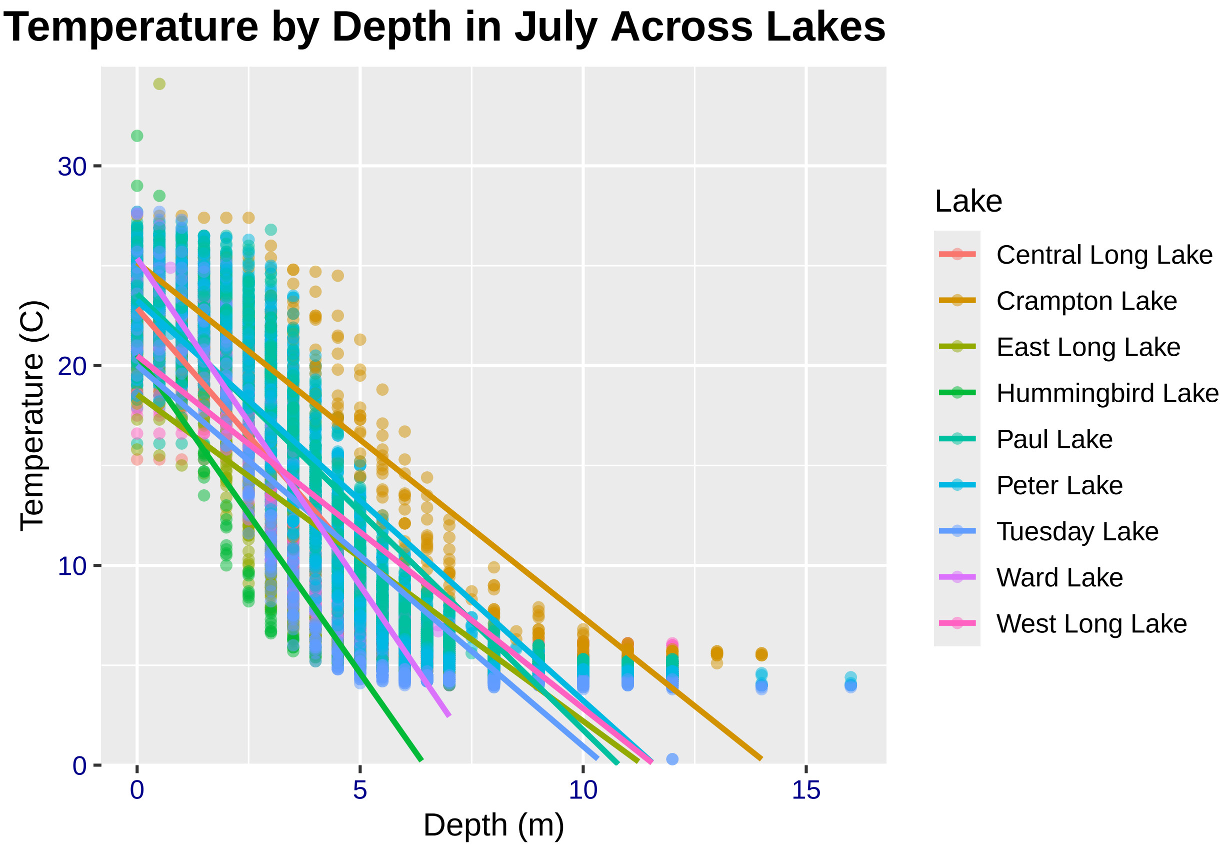
```r
#14.

lake.plot <- ggplot(NTL.filtered.lakename, aes(x = depth, y = temperature_C,
                                               color = lakename)) +
  geom_point(alpha = 0.5) +  # make points 50% transparent
  geom_smooth(method = "lm", se = FALSE) +  # add linear model line for each lake
  scale_y_continuous(limits = c(0, 35), expand = c(0, 0)) +  # set y-axis limits
  labs(
    title = "Temperature by Depth in July Across Lakes",
    x = "Depth (m)",
    y = "Temperature (C)",
    color = "Lake"
  ) +
  mytheme

lake.plot
```

```
## `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
tukey.lake <- TukeyHSD(lake_anova)
print(tukey.lake)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = NTL.filtered.lakename)
##
## $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
```

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: For lakes with the statistically same mean temperature, their p-value needs to be below 0.05 to reject the null hypothesis (which is that lake temperature is the same at all lakes). For all of the pairs including Peter Lake, this includes Paul Lake and Ward Lake. There are not any lakes with mean temperatures that are statistically distinct from all other lakes. Each lake has at least one other lake where the p-value is 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: You could perform a t-test. A t-test can be used to determine whether there is a significant difference between the means of two groups.

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?

```
# filter dataset
NTL.filtered.CramptonandWard <- NTL.raw %>%
  filter(month(sampledate) == 7)  %>%
  filter(lakename == c('Crampton Lake', 'Ward Lake')) %>% #month 7 = July
  select(lakename, year4, daynum, depth, temperature_C) %>%
  # Select specific columns
  drop_na() #drop entire cases with na's
```

```r
# two sample t test
NTL.twosample <- t.test(NTL.filtered.CramptonandWard$temperature_C ~
                          NTL.filtered.CramptonandWard$lakename)
NTL.twosample
```

```
##
##  Welch Two Sample t-test
##
## data:  NTL.filtered.CramptonandWard$temperature_C by NTL.filtered.CramptonandWard$lakename
## t = 1.4735, df = 90.058, p-value = 0.1441
## alternative hypothesis: true difference in means between group Crampton Lake and group Ward Lake is
## 95 percent confidence interval:
##  -0.5788109  3.9029455
## sample estimates:
## mean in group Crampton Lake     mean in group Ward Lake
##                    15.40437                    13.74231
```

Answer: The test provides a t-value of 1.4735 and a p-value of 0.1441. Because the p value is greater than 0.05, there is not sufficient evidence to conclude that the mean temperatures are different. In question 16, the results of the Tukey test also suggested that Crampton Lake and Ward Lake do not have a significant difference in mean temperatures. This conclusion aligns with the results of the t-test, which also found no significant difference. Both tests indicate that the mean temperatures for Crampton Lake and Ward Lake are statistically indistinguishable.