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

Student Name

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
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(agricolae)
library(here)
```

```
## here() starts at C:/Users/purec/Documents/Duke/Fall_2023/EDA/EDE_Fall2023
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
here()
```

```
## [1] "C:/Users/purec/Documents/Duke/Fall_2023/EDA/EDE_Fall2023"
```

```
LakeChemistry <- read.csv(file = here('Data','Raw',
                      'NTL-LTER_Lake_ChemistryPhysics_Raw.csv'),
                      stringsAsFactors = T)

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

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

```
#2

mytheme <- theme_gray()+
  theme(plot.title = element_text(size = 16, hjust= 0),
        axis.title = element_text(size = 13),
        legend.position = 'right')

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: There is no significant difference between temperatures recorded across depth measurements.

   mean depth 1 = mean depth 2 = mean depth 3 ... = mean depth n

   Ha: There are significant differences between temperatures recorded across depth measurements.

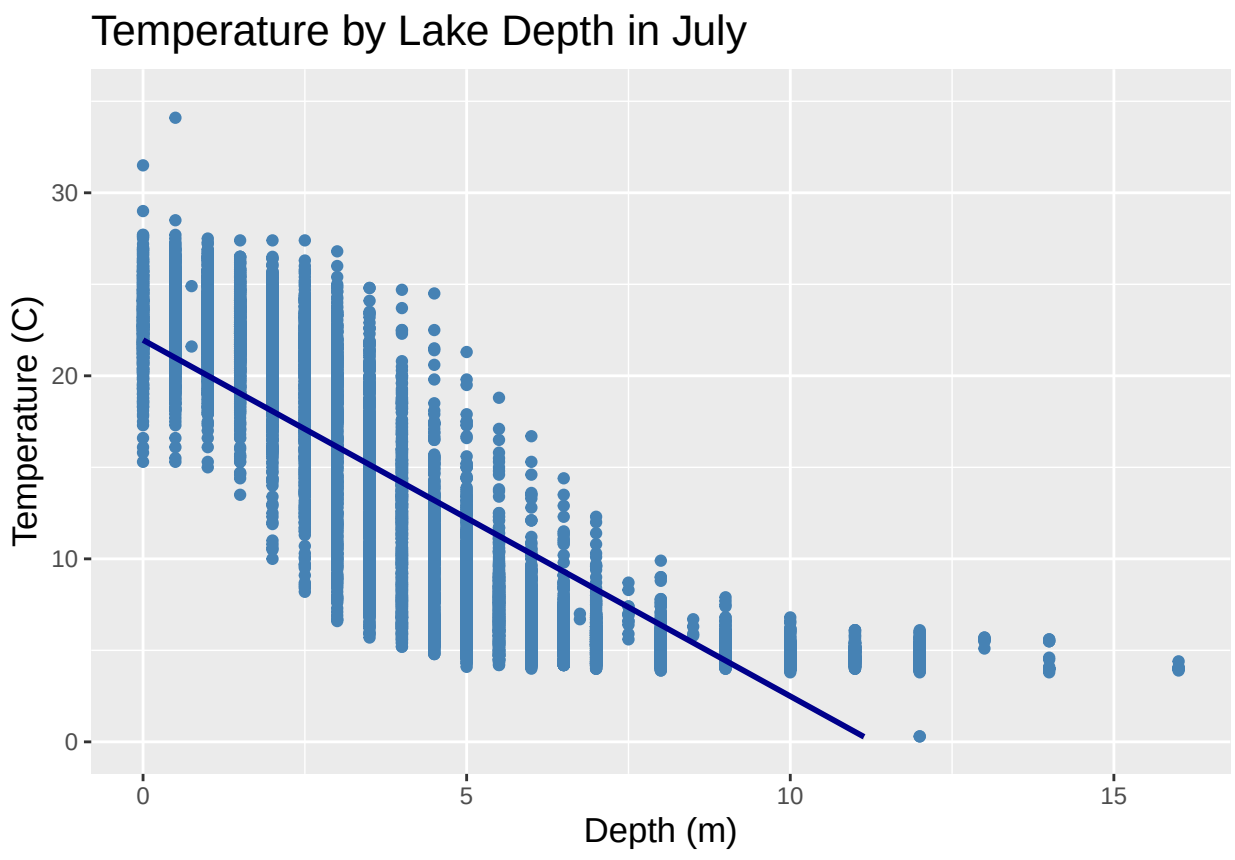
4. Wrangle your NTL-LTER dataset with a pipe function so that the records meet the following criteria:

- 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
LakeChemistry_July <- LakeChemistry %>%
  filter(month(sampledate)==7) %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  na.omit()

#5

Temp_by_Depth <- ggplot(LakeChemistry_July, aes(x=depth, y=temperature_C))+
  geom_point(color="steelblue")+
  geom_smooth(method="lm", color="darkblue")+
  ylim(0,35)+
  labs(x="Depth (m)",
       y="Temperature (C)",
       title="Temperature by Lake Depth in July")

Temp_by_Depth
```

```
## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 24 rows containing missing values (geom_smooth).
```

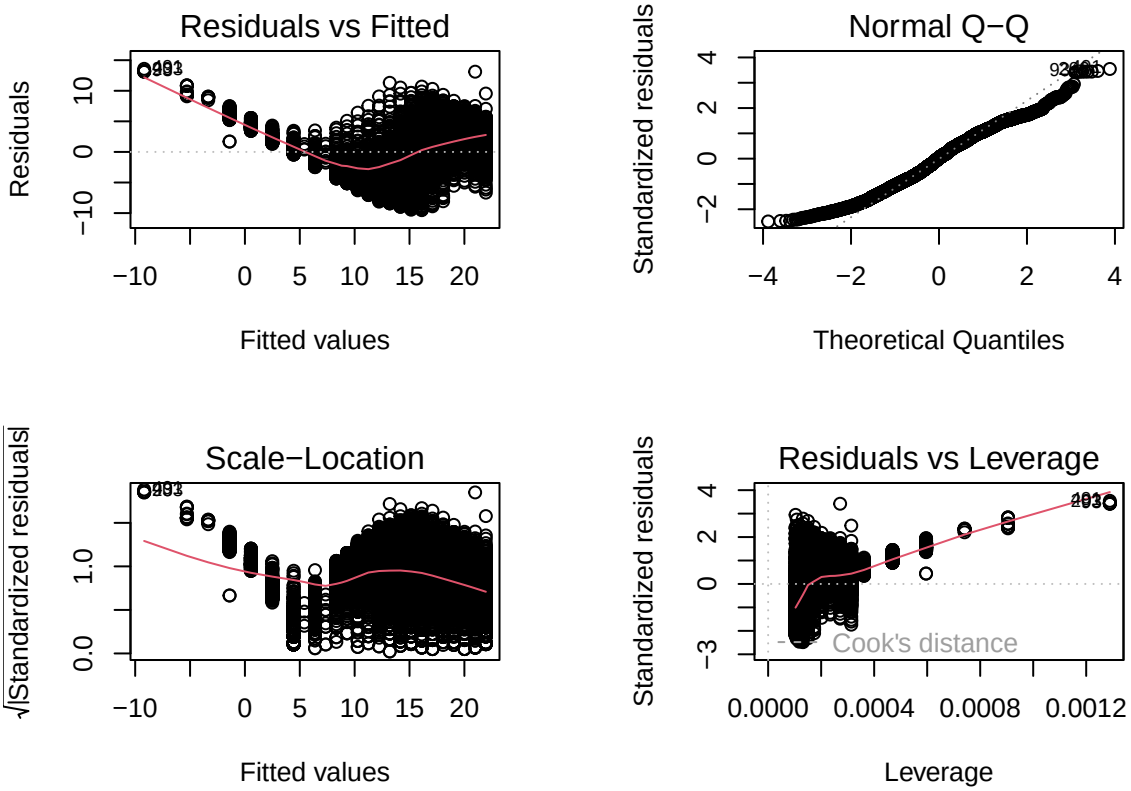


6. Interpret the figure. What does it suggest with regards to the response of temperature to depth? Do the distribution of points suggest about anything about the linearity of this trend?

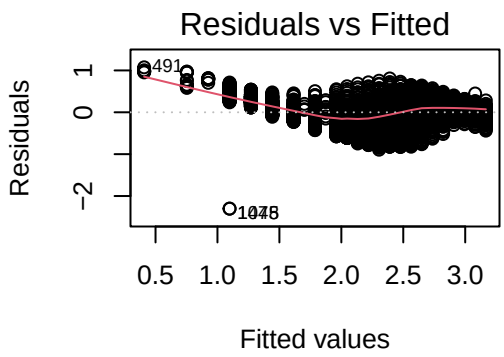
   Answer: Temperature decreases with increasing depth across both lakes. The distribution of points suggests perhaps a logarithmic relationship, as we can see that the points tend to converge at about 4C.

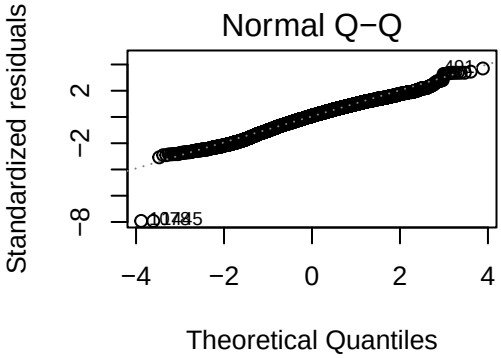
7. Perform a linear regression to test the relationship and display the results

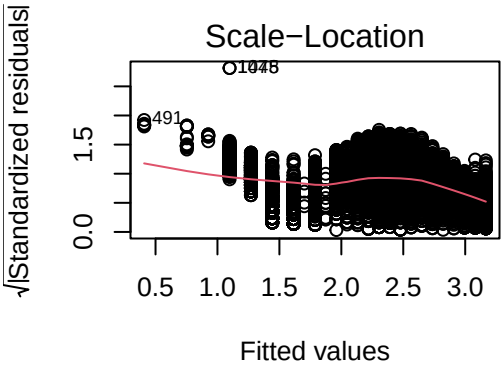
```
#7

#No logarithmics
temp_depth <- lm(data = LakeChemistry_July, temperature_C ~ depth)
par(mfrow = c(2,2), mar=c(4,4,4,4))
plot(temp_depth)
```

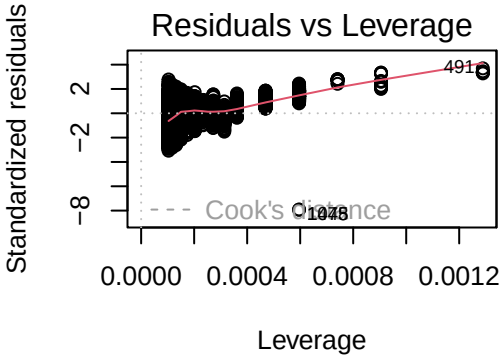



```
par(mfrow = c(1,1))

#With temperature logarithmic relationship
log_temp_depth <- lm(data = LakeChemistry_July, log(temperature_C) ~ depth)
par(mfrow = c(2,2), mar=c(4,4,4,4))
plot(log_temp_depth)
```










```
par(mfrow = c(1,1))

summary(log_temp_depth)
```

```
##
## Call:
## lm(formula = log(temperature_C) ~ depth, data = LakeChemistry_July)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.30179 -0.18012  0.02433  0.20886  1.07337
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.1665624  0.0051477   615.1   <2e-16 ***
## depth       -0.1723953  0.0008895  -193.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2907 on 9726 degrees of freedom
## Multiple R-squared:  0.7943, Adjusted R-squared:  0.7943
## F-statistic: 3.756e+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 following interpretation is for the log-linear model, with temperature log transformed. 79.43% of the variability in temperature can be explained by changes in depth, based on 9726 degrees of freedom. This is a very statistically significant result, as the p-value of this model approaches 0. For every 1m increase in depth, temperature is predicted to decrease by a factor of 0.172.

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
LakesAIC <- lm(data = LakeChemistry_July, log(temperature_C) ~ year4 + daynum + depth)
step(LakesAIC)
```

```
## Start:  AIC=-24131.28
## log(temperature_C) ~ year4 + daynum + depth
##
##          Df Sum of Sq    RSS    AIC
## - year4   1       0.2  813.7 -24131
## <none>                 813.5 -24131
## - daynum  1       8.0  821.5 -24039
## - depth   1    3173.1 3986.6  -8672
##
## Step:  AIC=-24131.33
## log(temperature_C) ~ daynum + depth
##
##          Df Sum of Sq    RSS      AIC
## <none>                 813.7 -24131.3
## - daynum  1         8  821.6 -24038.5
## - depth   1      3173 3986.7  -8673.9

##
## Call:
## lm(formula = log(temperature_C) ~ daynum + depth, data = LakeChemistry_July)
##
## Coefficients:
## (Intercept)       daynum        depth
##    2.536006     0.003193    -0.172387
```

```
#10
summary(LakesAIC)
```

```
##
## Call:
## lm(formula = log(temperature_C) ~ year4 + daynum + depth, data = LakeChemistry_July)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.28889 -0.18120  0.02367  0.20917  1.05894
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  1.6268138  0.6539728    2.488   0.0129 *
## year4        0.0004551  0.0003257    1.397   0.1624
## daynum       0.0031905  0.0003271    9.754   <2e-16 ***
## depth       -0.1724004  0.0008852 -194.755   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2892 on 9724 degrees of freedom
## Multiple R-squared:  0.7964, Adjusted R-squared:  0.7963
## F-statistic: 1.268e+04 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 of explanatory variables is all three– year4, daynum, and depth. We know this because after running step(LakesAIC), the lowest AIC value resulted when none of those three variables were removed from the model. This model explains 79.64% of the observed variance, a 0.21% increase from our previous model. This is a very slight improvement, but still worth conducting.

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
LakeTemps_aov <- aov(data = LakeChemistry_July, temperature_C ~ lakename)
summary(LakeTemps_aov)
```

```
##               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
LakeTemps_lm <- lm(data = LakeChemistry_July, temperature_C ~ lakename)
summary(LakeTemps_lm)
```

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

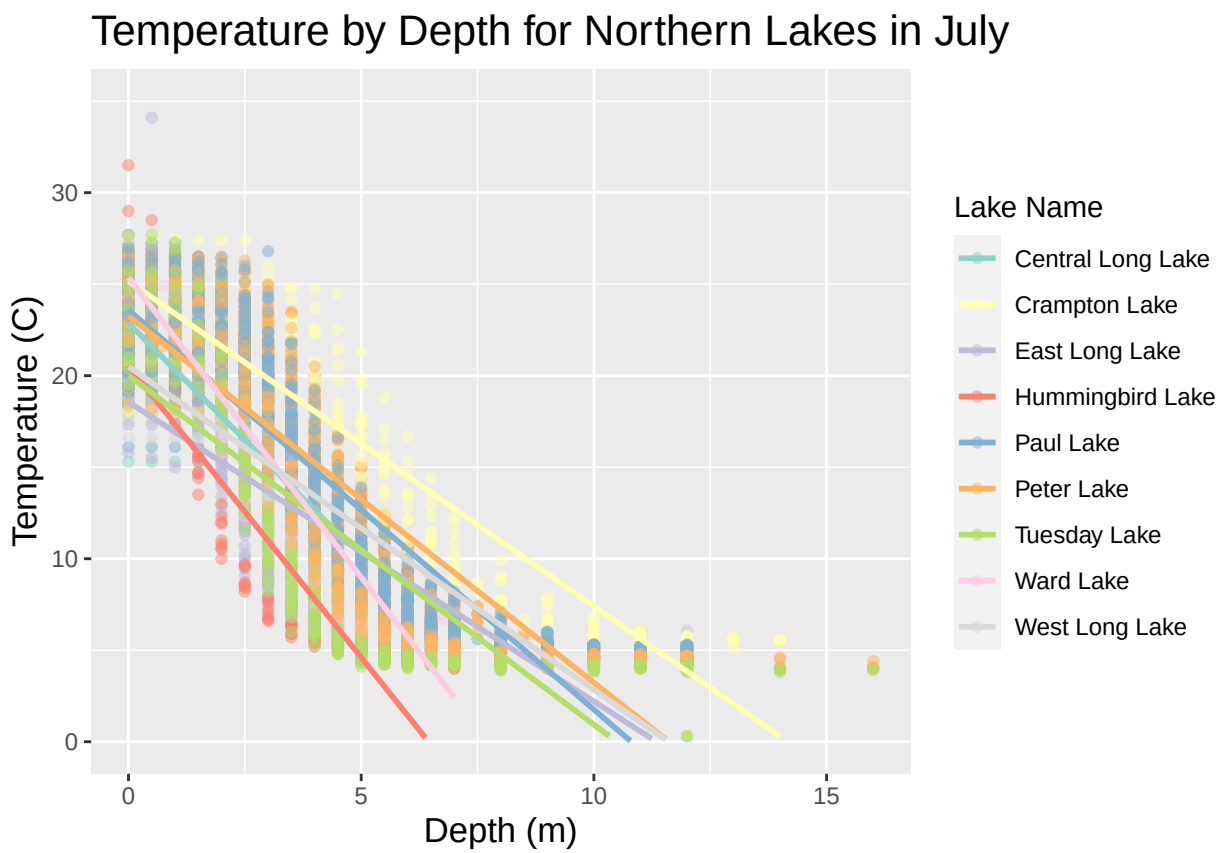
    Answer: Yes, there is a significant difference in mean temperature among the lakes. In the ANOVA summary, the overall p-value for lakename is less than 0.001, which is a very significant result. In the linear model results, every lake had a p-value less than 0.05.

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.

Temp_by_Depth_Lakes <- ggplot(LakeChemistry_July,
                              aes(x=depth, y=temperature_C, color=lakename))+
  geom_point(alpha=0.5)+
  geom_smooth(method="lm", se=FALSE)+
  scale_color_brewer(palette='Set3')+
  ylim(0,35)+
  labs(x="Depth (m)",
       y="Temperature (C)",
       color='Lake Name',
       title="Temperature by Depth for Northern Lakes in July")

Temp_by_Depth_Lakes
```

```
## `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
LakeTemp.groups <- HSD.test(LakeTemps_aov, "lakename", group = TRUE)
LakeTemp.groups
```

```
## $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: Statistically, Paul Lake and Ward Lake have the same mean temperature as Peter Lake, indicated by the fact that they all have the group c label. No lake has a mean temperature statistically distinct from the others, as all have overlap with at least one other 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: We could use a t-test since we are only looking at two distinct 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?

```
Crampton_Ward_July <- LakeChemistry_July %>%
  subset(lakename == 'Crampton Lake' | lakename == 'Ward Lake')

CW.twosample <- t.test(Crampton_Ward_July$temperature_C ~ Crampton_Ward_July$lakename)
CW.twosample
```

```
##
##  Welch Two Sample t-test
##
## data:  Crampton_Ward_July$temperature_C by Crampton_Ward_July$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: The test says that the July temperatures for Crampton and Ward Lakes are not statistically different. The results say that the p-value is 0.2649, which is greater than 0.05 which we typically use as our alpha. This means that we do not reject the null hypothesis that the temperature means are equal, which aligns with question 16 results because the Tukey HSD results grouped Crampton and Ward Lakes in group b to indicate that they have statistically similar temperature values.