Assignment 6: GLMs (Linear Regressios, ANOVA, & t-tests)

Joanna Huertas

Spring 2023

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on generalized linear models.

Directions

1. Rename this file `<FirstLast>_A06_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.

```r
#1
library(here)
```

```
## here() starts at C:/Users/joann/Documents/EDA-Spring2023
```

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(agricolae)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(lubridate)
```

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

```
here()
```

```
## [1] "C:/Users/joann/Documents/EDA-Spring2023"
```

```
NTL_LTER <- read.csv(here("Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv"), stringsAsFactors = TRUE)

# convert the "Date" column to a date object
NTL_LTER$sampledate <- as.Date(NTL_LTER$sampledat, format = "%m/%d/%y")

#2
# Set theme
mytheme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
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: Yes, as we can see later in the graphic, there is a negative correlation between temperature and depth. H0: depth across all lakes has no correlation with the temperature Ha: depth across all lakes has correlation with the temperature

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

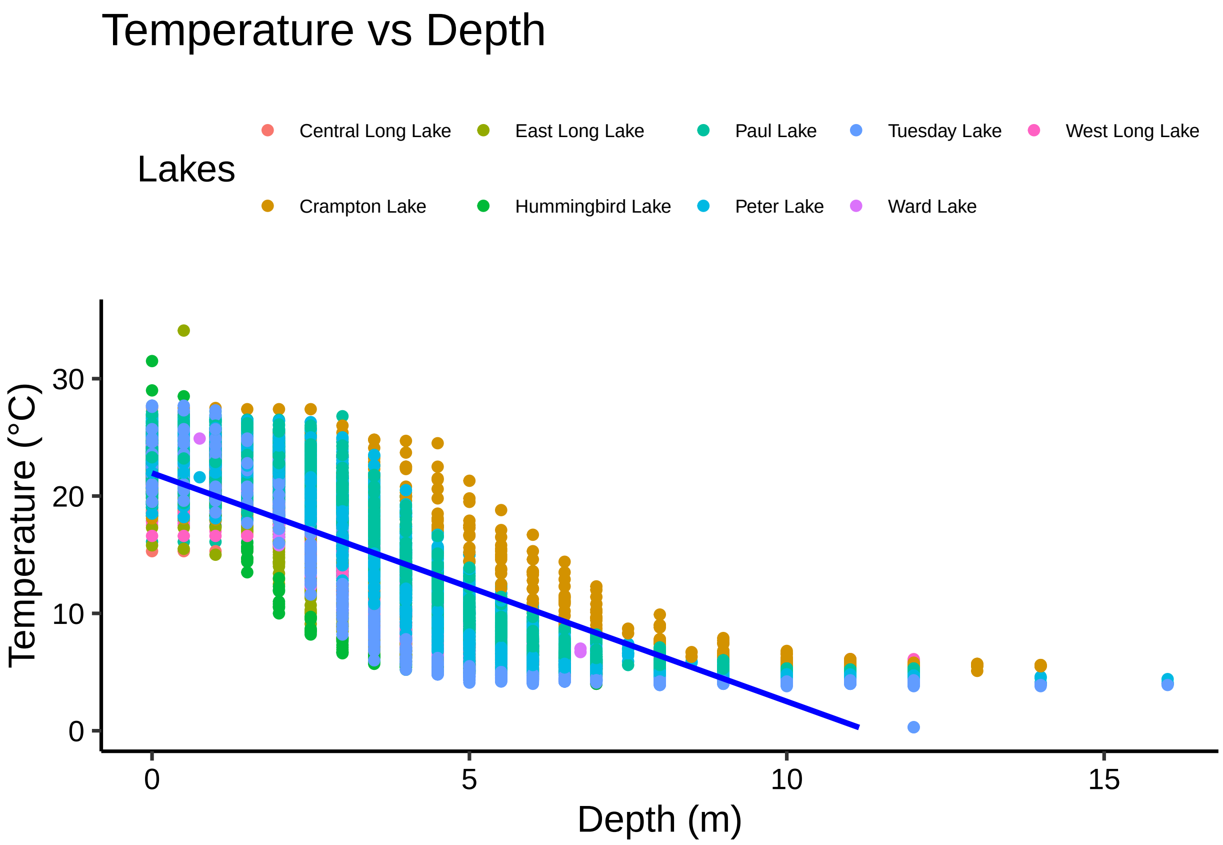
- Only dates in July.
- Only the columns: `lakename`, `year4`, `daynum`, `depth`, `temperature_C`
- Only complete cases (i.e., remove NAs)

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_LTER.wrangle <-
  NTL_LTER %>%
  filter(month(sampledate)==7)%>%
  select(lakename, year4, daynum, depth, temperature_C)%>%
  drop_na()


#5
scatterplot <-
  ggplot(NTL_LTER.wrangle, aes(x=depth, y=temperature_C, color = lakename))+
  geom_point()+
  geom_smooth(method="lm", se = FALSE, color = "blue") +
  #xlim(0, 125) +
  ylim(0, 35)+
  labs(title= "Temperature vs Depth", x= "Depth (m)", y= "Temperature (°C)", color="Lakes")+
   theme(legend.text = element_text(size = 7), legend.position = "top",
         legend.justification = "right", legend.key.width = unit(10, "pt"),
         legend.key.height = unit(1, "cm"))

print(scatterplot)
```

```
## `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: In this case, the scatterplot suggests that temperature is a function of depth, with temperature decreasing as depth increases, indicating a strong negative relationship.

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

```
#7

temp.depth.regression <- lm(data = NTL_LTER.wrangle, depth ~ temperature_C)
summary(temp.depth.regression)
```

```
##
## Call:
## lm(formula = depth ~ temperature_C, data = NTL_LTER.wrangle)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0685 -1.1065 -0.2334  0.9668  8.0964
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.573728   0.033803   283.2   <2e-16 ***
## temperature_C -0.379578   0.002289  -165.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.694 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 analysis indicates that changes in depth explain 73.87% of the variability in temperature (notice that R-squared is close to 1). This finding is based on 1 degree of freedom and 9726 degrees of freedom for the residuals, with a residual standard error of 1.694. The F-statistic for the regression is 2.75e+04 with a p-value of < 2.2e-16 (p<alpha), indicating that the regression is statistically significant. Change in temperature = -0.379578 x 1 = -0.379578°C Therefore, for every 1 meter increase in depth, temperature is predicted to decrease by approximately 0.38°C, according to the linear regression analysis.

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

cor.test(NTL_LTER.wrangle$depth, NTL_LTER.wrangle$temperature_C)
```

```
##
##  Pearson's product-moment correlation
##
## data:  NTL_LTER.wrangle$depth and NTL_LTER.wrangle$temperature_C
## t = -165.83, df = 9726, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8646036 -0.8542169
## sample estimates:
##        cor
## -0.8594989
```

```
TPAIC.1 <- lm(data = NTL_LTER.wrangle, temperature_C ~ year4 + daynum + depth)

#Choose a model by AIC in a Stepwise Algorithm
step(TPAIC.1)
```

```
## 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_LTER.wrangle)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##    -8.57556      0.01134      0.03978     -1.94644
```

```
#10
TPmodel <- lm(data = NTL_LTER.wrangle, temperature_C ~ year4 + daynum +
              depth)
summary(TPmodel)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL_LTER.wrangle)
##
## 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 AIC method suggests using all three explanatory variables, “year4”, “daynum”, and “depth”, to predict temperature in the multiple regression. The model explains 100% - (141687 / 26066) = 81.54% of the observed variance in the data.The initial AIC value was 26065.53, and the model including all three explanatory variables had an AIC of 26066, which is only slightly higher. This suggests that the additional explanatory variables have not significantly improved the model’s fit.

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
# Wrangle the data
NTL_LTER.temperatures <- NTL_LTER.wrangle %>%
  group_by(year4, daynum, lakename) %>%
  summarise(temperature_C = mean(temperature_C))
```

```
## `summarise()` has grouped output by 'year4', 'daynum'. You can override using
## the `.groups` argument.
```

```
summary(NTL_LTER.temperatures)
```

```
##      year4          daynum                   lakename   temperature_C
##  Min.   :1984   Min.   :182.0   Paul Lake      :150   Min.   : 8.645
##  1st Qu.:1991   1st Qu.:190.0   Peter Lake     :150   1st Qu.:11.190
##  Median :1997   Median :198.0   Tuesday Lake   : 86   Median :12.700
##  Mean   :1999   Mean   :197.5   West Long Lake: 52   Mean   :12.865
##  3rd Qu.:2006   3rd Qu.:205.0   East Long Lake: 49   3rd Qu.:14.150
##  Max.   :2016   Max.   :213.0   Crampton Lake : 15   Max.   :21.022
##                                 (Other)       : 31
```

```
#check all value of lakename
summary(NTL_LTER.temperatures$lakename)
```

```
## Central Long Lake     Crampton Lake    East Long Lake  Hummingbird Lake
##                14                15                49                 9
##         Paul Lake        Peter Lake      Tuesday Lake         Ward Lake
##               150               150                86                 8
##    West Long Lake
##                52
```

```
# Format ANOVA as aov
NTL_LTER.temperatures.anova <- aov(data = NTL_LTER.temperatures, temperature_C ~ lakename)
summary(NTL_LTER.temperatures.anova)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## lakename       8   1333  166.60   79.69 <2e-16 ***
## Residuals    524   1096    2.09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

13. Is there a significant difference in mean temperature among the lakes? Report your findings.

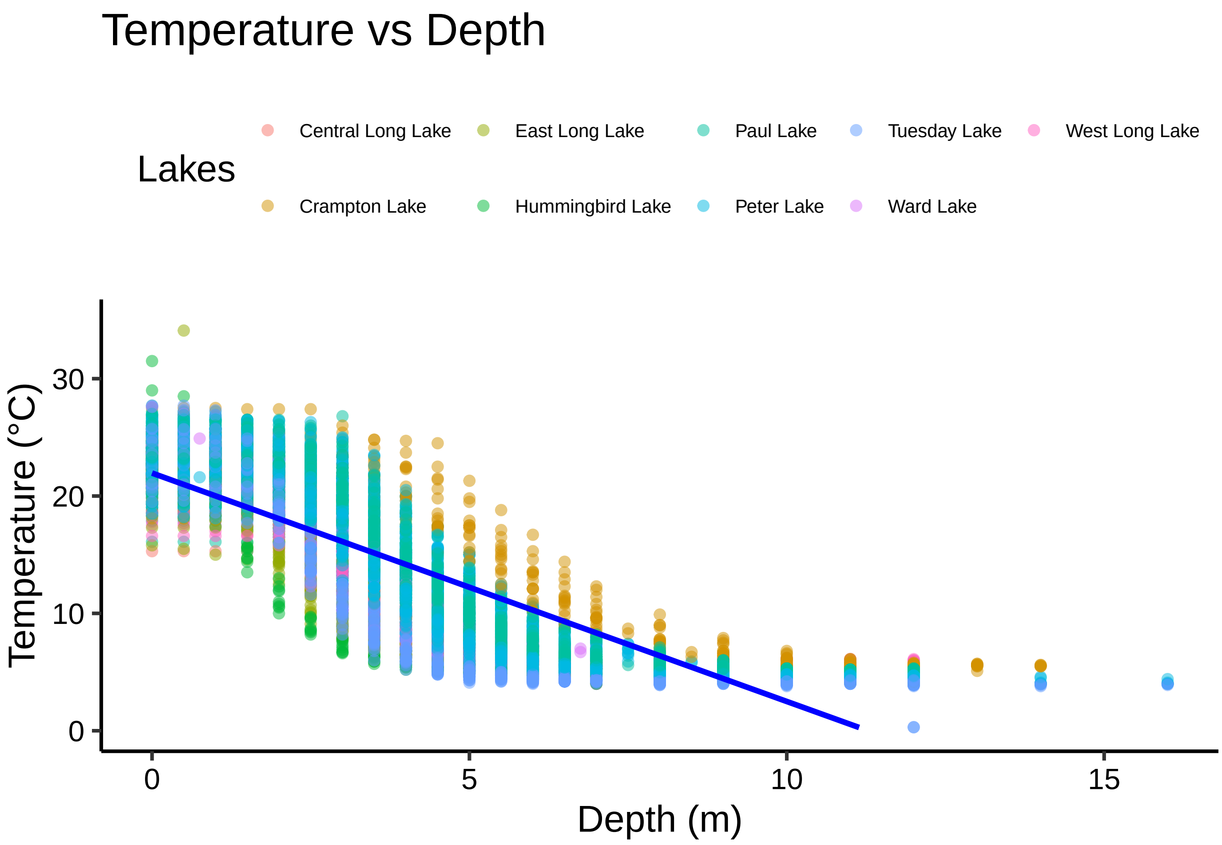
    Answer: Yes. The P value is less than 0.05, then we reject the null hypothesis (the mean is the same across all different lakes), this the means are not all the same.

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.

scatterplot2 <-
ggplot(NTL_LTER.wrangle, aes(x=depth, y=temperature_C, color = lakename))+
  geom_point(alpha = 0.5)+
geom_smooth(method="lm", color="blue", se=FALSE)+
  #xlim(0, 125) +
  ylim(0, 35)+
  labs(fill="", title= "Temperature vs Depth", x= "Depth (m)", y= "Temperature (°C)", color="Lakes")+
   theme(legend.text = element_text(size = 7), legend.position = "top",
         legend.justification = "right", legend.key.width = unit(10, "pt"),
         legend.key.height = unit(1, "cm"))

print(scatterplot2)
```

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

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



15. Use the Tukey's HSD test to determine which lakes have different means.

```
#15
# Post-hoc test
# TukeyHSD() computes Tukey Honest Significant Differences
TukeyHSD(NTL_LTER.temperatures.anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = NTL_LTER.temperatures)
##
## $lakename
##                                          diff         lwr         upr     p adj
## Crampton Lake-Central Long Lake    -2.3175579 -3.99122964 -0.64388624 0.0006463
## East Long Lake-Central Long Lake   -7.4019316 -8.76679490 -6.03706840 0.0000000
## Hummingbird Lake-Central Long Lake -6.8654742 -8.78971561 -4.94123270 0.0000000
## Paul Lake-Central Long Lake        -3.8099221 -5.06853792 -2.55130621 0.0000000
## Peter Lake-Central Long Lake       -4.2465818 -5.50519766 -2.98796595 0.0000000
## Tuesday Lake-Central Long Lake     -6.4967571 -7.79473610 -5.19877803 0.0000000
## Ward Lake-Central Long Lake        -3.2574254 -5.25352980 -1.26132101 0.0000182
## West Long Lake-Central Long Lake   -6.1034132 -7.45949914 -4.74732734 0.0000000
## East Long Lake-Crampton Lake       -5.0843737 -6.41338062 -3.75536680 0.0000000
## Hummingbird Lake-Crampton Lake     -4.5479162 -6.44689301 -2.64893942 0.0000000
## Paul Lake-Crampton Lake            -1.4923641 -2.71200406 -0.27272419 0.0048437
## Peter Lake-Crampton Lake           -1.9290239 -3.14866380 -0.70938393 0.0000394
## Tuesday Lake-Crampton Lake         -4.1791991 -5.43942024 -2.91897800 0.0000000
## Ward Lake-Crampton Lake            -0.9398675 -2.91162821  1.03189328 0.8624463
## West Long Lake-Crampton Lake       -3.7858553 -5.10584646 -2.46586414 0.0000000
## Hummingbird Lake-East Long Lake     0.5364575 -1.09687884  2.16979383 0.9835941
## Paul Lake-East Long Lake            3.5920096  2.85093181  4.33308736 0.0000000
## Peter Lake-East Long Lake           3.1553498  2.41427207  3.89642762 0.0000000
## Tuesday Lake-East Long Lake         0.9051746  0.09905302  1.71129615 0.0148765
## Ward Lake-East Long Lake            4.1445062  2.42709099  5.86192150 0.0000000
## West Long Lake-East Long Lake       1.2985184  0.40182930  2.19520753 0.0002729
## Paul Lake-Hummingbird Lake          3.0555521  1.50989696  4.60120722 0.0000001
## Peter Lake-Hummingbird Lake         2.6188923  1.07323722  4.16454748 0.0000068
## Tuesday Lake-Hummingbird Lake       0.3687171 -1.20915663  1.94659082 0.9983857
## Ward Lake-Hummingbird Lake          3.6080488  1.41958612  5.79651138 0.0000140
## West Long Lake-Hummingbird Lake     0.7620609 -0.86394796  2.38806980 0.8734535
## Peter Lake-Paul Lake               -0.4366597 -0.95671595  0.08339647 0.1826605
## Tuesday Lake-Paul Lake             -2.6868350 -3.29600999 -2.07766000 0.0000000
## Ward Lake-Paul Lake                 0.5524967 -1.08175465  2.18674797 0.9802906
## West Long Lake-Paul Lake           -2.2934912 -3.01827635 -1.56870599 0.0000000
## Tuesday Lake-Peter Lake            -2.2501753 -2.85935025 -1.64100026 0.0000000
## Ward Lake-Peter Lake                0.9891564 -0.64509491  2.62340771 0.6242681
## West Long Lake-Peter Lake          -1.8568314 -2.58161661 -1.13204625 0.0000000
## Ward Lake-Tuesday Lake              3.2393317  1.57457550  4.90408782 0.0000001
## West Long Lake-Tuesday Lake         0.3933438 -0.39782573  1.18451338 0.8317994
## West Long Lake-Ward Lake           -2.8459878 -4.55643586 -1.13553981 0.0000111
```

```
# Extract groupings for pairwise relationships
NTL_LTER.temperatures.groups <- HSD.test(NTL_LTER.temperatures.anova, "lakename", group = TRUE)
NTL_LTER.temperatures.groups
```

```
## $statistics
##   MSerror  Df     Mean       CV
##   2.09074 524 12.86544 11.23894
##
## $parameters
##    test   name.t ntr StudentizedRange alpha
##   Tukey lakename   9            4.405  0.05
##
## $means
##                     temperature_C       std   r       Min      Max      Q25
## Central Long Lake      17.68698 1.9058275  14 14.544444 21.02222 16.66389
## Crampton Lake          15.36943 1.5138376  15 12.909091 18.10476 14.47045
## East Long Lake         10.28505 1.0519538  49  8.645000 12.70000  9.49000
## Hummingbird Lake       10.82151 0.7962931   9  9.146667 11.88462 10.45385
## Paul Lake              13.87706 1.2358897 150 10.900000 17.80625 12.97778
## Peter Lake             13.44040 1.8099936 150 10.165000 19.54000 12.14708
## Tuesday Lake           11.19023 1.4616471  86  8.935000 18.34545 10.43281
## Ward Lake              14.42956 1.5572801   8 12.360000 16.10714 13.31141
## West Long Lake         11.58357 0.9263405  52  9.745000 13.49500 10.99250
##                          Q50      Q75
## Central Long Lake 17.61889 18.71389
## Crampton Lake     15.11818 16.25000
## East Long Lake    10.08500 10.98500
## Hummingbird Lake  10.97692 11.16154
## Paul Lake          13.85000 14.57500
## Peter Lake         13.24500 14.55319
## Tuesday Lake       10.91111 11.70937
## Ward Lake          14.54282 15.77156
## West Long Lake     11.58000 12.25500
##
## $comparison
## NULL
##
## $groups
##                    temperature_C groups
## Central Long Lake       17.68698      a
## Crampton Lake           15.36943      b
## Ward Lake               14.42956     bc
## Paul Lake               13.87706      c
## Peter Lake              13.44040      c
## West Long Lake          11.58357      d
## Tuesday Lake            11.19023      d
## Hummingbird Lake        10.82151     de
## East Long Lake          10.28505      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 has the same mean temperature as Peter Lake. Central Long Lake has a mean temperature that is statistically distinct from all the other lakes.

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: If we are only interested in comparing the mean temperatures of Peter Lake and Paul Lake, we could use a two-sample t-test to determine whether their means are statistically different.

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
NTL_LTER.Crampton.Ward.Lakes <- NTL_LTER.wrangle %>%
  filter(lakename== "Crampton Lake" | lakename== "Ward Lake")%>%
  group_by(year4, daynum, lakename) %>%
  summarise(temperature_C = mean(temperature_C))
```

```
## `summarise()` has grouped output by 'year4', 'daynum'. You can override using
## the `.groups` argument.
```

```
summary(NTL_LTER.Crampton.Ward.Lakes)
```

```
##      year4          daynum                   lakename  temperature_C
##  Min.   :1999   Min.   :183.0   Crampton Lake    :15   Min.   :12.36
##  1st Qu.:2004   1st Qu.:189.5   Ward Lake        : 8   1st Qu.:13.87
##  Median :2005   Median :197.0   Central Long Lake: 0   Median :15.12
##  Mean   :2006   Mean   :196.8   East Long Lake   : 0   Mean   :15.04
##  3rd Qu.:2010   3rd Qu.:203.0   Hummingbird Lake : 0   3rd Qu.:16.06
##  Max.   :2012   Max.   :211.0   Paul Lake        : 0   Max.   :18.10
##                                 (Other)          : 0
```

```
#check all value of lakename
summary(NTL_LTER.Crampton.Ward.Lakes$lakename)
```

```
## Central Long Lake     Crampton Lake    East Long Lake  Hummingbird Lake
##                 0                15                 0                 0
##         Paul Lake        Peter Lake      Tuesday Lake         Ward Lake
##                 0                 0                 0                 8
##    West Long Lake
##                 0
```

```
#Format as a t-test
#EPAair$Ozone will be our continuous dependent variable
#EPAair$Year will be our categorical variable with two levels (2018 and 2019)
Crampton.Ward.twosample <- t.test(NTL_LTER.Crampton.Ward.Lakes$temperature_C ~
                                    NTL_LTER.Crampton.Ward.Lakes$lakename)
Crampton.Ward.twosample
```

```
##
##  Welch Two Sample t-test
##
## data:  NTL_LTER.Crampton.Ward.Lakes$temperature_C by NTL_LTER.Crampton.Ward.Lakes$lakename
## t = 1.3919, df = 14.05, p-value = 0.1856
## alternative hypothesis: true difference in means between group Crampton Lake and group Ward Lake is
## 95 percent confidence interval:
##  -0.5078534  2.3875883
## sample estimates:
## mean in group Crampton Lake     mean in group Ward Lake
##                    15.36943                    14.42956
```

```
# Format as a GLM
Crampton.Ward.twosample2 <- lm(NTL_LTER.Crampton.Ward.Lakes$temperature_C ~
                                 NTL_LTER.Crampton.Ward.Lakes$lakename)
summary(Crampton.Ward.twosample2)
```

```
##
## Call:
## lm(formula = NTL_LTER.Crampton.Ward.Lakes$temperature_C ~ NTL_LTER.Crampton.Ward.Lakes$lakename)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4603 -1.0387 -0.2512  1.1638  2.7353
##
## Coefficients:
##                                               Estimate Std. Error t value
## (Intercept)                                    15.3694     0.3946  38.945
## NTL_LTER.Crampton.Ward.Lakes$lakenameWard Lake  -0.9399     0.6692  -1.405
##                                               Pr(>|t|)
## (Intercept)                                     <2e-16 ***
## NTL_LTER.Crampton.Ward.Lakes$lakenameWard Lake    0.175
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.528 on 21 degrees of freedom
## Multiple R-squared:  0.08588,    Adjusted R-squared:  0.04235
## F-statistic: 1.973 on 1 and 21 DF,  p-value: 0.1748
```

Answer: Our p-value is higher than 0.05 so we don't reject the null hypothesis, therefore, the two means (Crampton Lake and Ward Lake temperatures) are the same.My answer in part 16 says that Crampton Lake and Ward Lake have similar means but not the same, because Crampton Lake has the group b and Ward Lake has the group bc, so they only share b.