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

Shubhangi Gupta

Spring 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
#Checking working directory
getwd()
```

```
## [1] "/home/guest/RStudio Project Folder/EDA_Spring2024"
```

```
#loading packages
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## 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(agricolae)
library(ggplot2)
library(lubridate)
library(here)
```

```
## here() starts at /home/guest/RStudio Project Folder/EDA_Spring2024
```

```r
#Importing the dataset
data_raw <- read.csv("Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv", stringsAsFactors = TRUE)

#Changing the date column to a date object
data_raw$sampledate <- mdy(data_raw$sampledate)

#Checking dataset
glimpse(data_raw)
```

```
## Rows: 38,614
## Columns: 11
## $ lakeid          <fct> L, L, L, L, L, L, L, L, L, L, L, L, L, L, L, L, L, R, ~
## $ lakename        <fct> Paul Lake, Paul Lake, Paul Lake, Paul Lake, Paul Lake,~
## $ year4           <int> 1984, 1984, 1984, 1984, 1984, 1984, 1984, 1984, 1984, ~
## $ daynum          <int> 148, 148, 148, 148, 148, 148, 148, 148, 148, 148, 148,~
## $ sampledate      <date> 1984-05-27, 1984-05-27, 1984-05-27, 1984-05-27, 1984--~
## $ depth           <dbl> 0.00, 0.25, 0.50, 0.75, 1.00, 1.50, 2.00, 3.00, 4.00, ~
## $ temperature_C   <dbl> 14.5, NA, NA, NA, 14.5, NA, 14.2, 11.0, 7.0, 6.1, 5.5,~
## $ dissolvedOxygen <dbl> 9.5, NA, NA, NA, 8.8, NA, 8.6, 11.5, 11.9, 2.5, 1.6, 0~
## $ irradianceWater <dbl> 1750.0, 1550.0, 1150.0, 975.0, 870.0, 610.0, 420.0, 22~
## $ irradianceDeck  <dbl> 1620, 1620, 1620, 1620, 1620, 1620, 1620, 1620, 1620, ~
## $ comments        <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
```

```r
#2
mytheme <- theme_classic(base_size = 14)+
  theme(axis.text = element_text (color = "black"), 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: mean temperature in July does not change with depth Ha: mean temperature in July changes with 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)

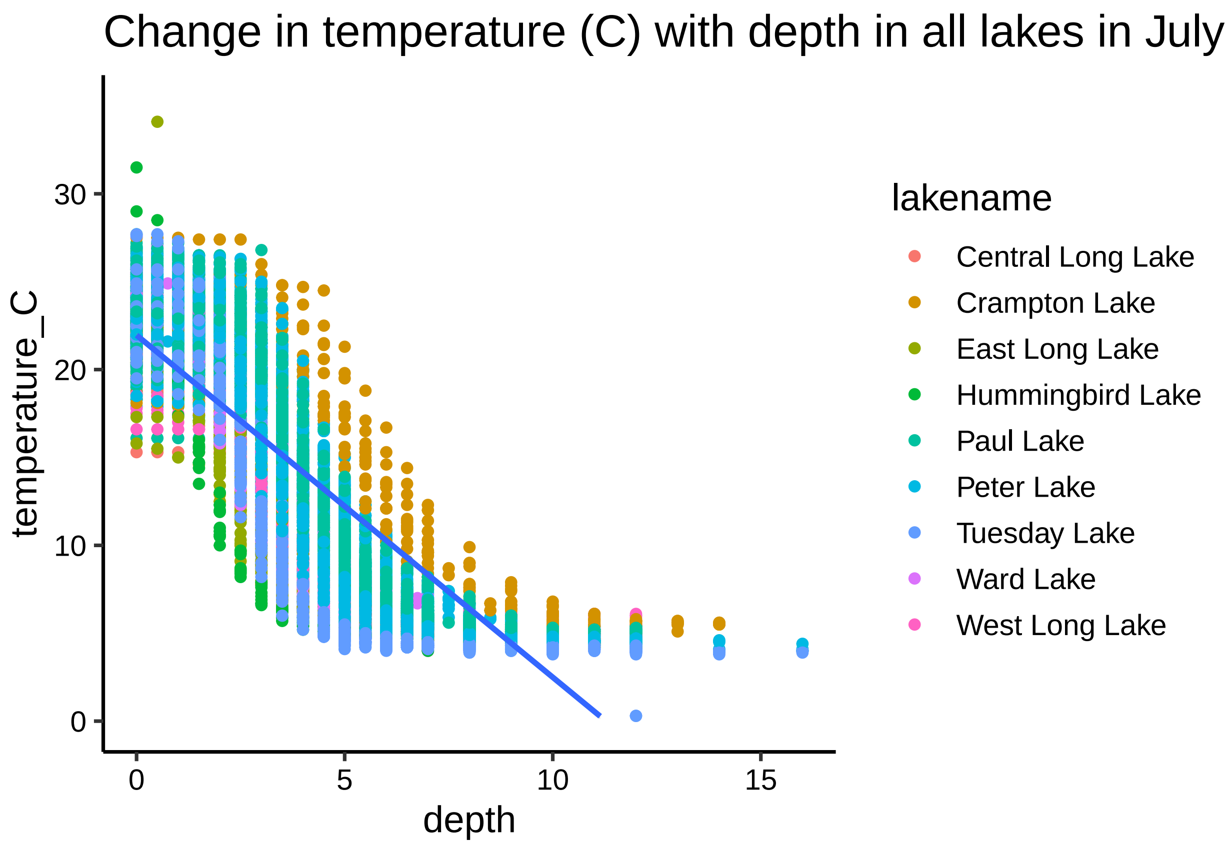
5. Visualize the relationship among the two continuous variables with a scatter plot of temperature by depth. Add a smoothed line showing the linear model, and limit temperature values from 0 to 35 °C. Make this plot look pretty and easy to read.

```r
#4
data_processed <- data_raw %>%
  mutate(month = month(sampledate)) %>%
  filter(month == "7") %>%
  select(lakename, year4, daynum, depth, temperature_C)%>%
  drop_na(c("depth", "temperature_C"))

glimpse(data_processed)
```

```
## Rows: 9,728
## Columns: 5
## $ lakename      <fct> Paul Lake, Paul Lake, Paul Lake, Paul Lake, Paul Lake, P~
## $ year4         <int> 1984, 1984, 1984, 1984, 1984, 1984, 1984, 1984, 1984, 19~
## $ daynum        <int> 183, 183, 183, 183, 183, 183, 183, 183, 183, 183, 183, 1~
## $ depth         <dbl> 0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 5.0, 6.0, 7~
## $ temperature_C <dbl> 22.8, 22.9, 22.8, 22.7, 21.7, 20.3, 18.2, 14.8, 12.3, 8.~
```

```r
#5
ggplot(data_processed, aes(x=depth, y=temperature_C))+
  geom_point(aes(color = lakename))+
  geom_smooth(method="lm")+
  ylim(0,35)+
  ggtitle("Change in temperature (C) with depth in all lakes in July")
```

```
## `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: Yes, these two variables seem to be inversely proportional, ie an increase in depth leads to a decrease in temperature.

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

```r
#7

tempvsdepth_lm <- lm(data_processed, formula = temperature_C ~ depth)
summary(tempvsdepth_lm)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = data_processed)
##
## 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 slope of temperature vs depth is -1.94 indicating that for every 1m increase in depth, there is a 1.94 deg C decrease in temperature of the lake. The p-value of both the coefficients, as well as of the overall F-statistic is < 0.05, indicating that this relationship between them is statistically significant. The R-squared value is 0.7387 indicating that 73% of the variability in temperature change is explained by the change in depth of the lake, and this is based on 9726 degrees of freedom.

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
temp_AIC <- lm(data=data_processed, temperature_C ~ year4 + daynum + depth)
print(step(temp_AIC))
```

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

```r
print(cor(data_processed[,c(2:5)]))
```

```
##                     year4        daynum         depth temperature_C
## year4         1.000000000  0.0048603276  0.0105584225    0.00477053
## daynum        0.004860328  1.0000000000 -0.0009266367    0.04840330
## depth         0.010558422 -0.0009266367  1.0000000000   -0.85949893
## temperature_C 0.004770530  0.0484033019 -0.8594989332    1.00000000
```

```r
#10
summary(temp_AIC)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = data_processed)
##
## 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: All three of the suggested variables seem to be statistically signficantly related to temperature, as the AIC increased when either of the three were removed. Thus I retained all three and the results of the multi linear regression regressing temperature on the three explanatory variables of year, day and depth led to a statistically significant output with an overall p-value much lesser than 0.05 (of the F statistic) and R-squared value of 0.74 meaning the three explanatory variables explain 74% of the variability in temperature based on 9724 degrees of freedom. However, the strongest correlation was with depth with a slope of -1.94 compared to the much lesser 0.04 and 0.01 with daynum and year respectively. This is validated by the correlation matrix that gives a -0.86 correlation between temperature and depth, but only 0.0047 and 0.048 with year and daynum respectively.

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
#anova model
tempvslakename_anova <- aov(data = data_processed, temperature_C ~ lakename)
summary(tempvslakename_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
```

```
#GLM model
tempvsdepth_anova_glm <- lm(data = data_processed, formula = temperature_C ~ lakename)
summary(tempvsdepth_anova_glm)
```

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

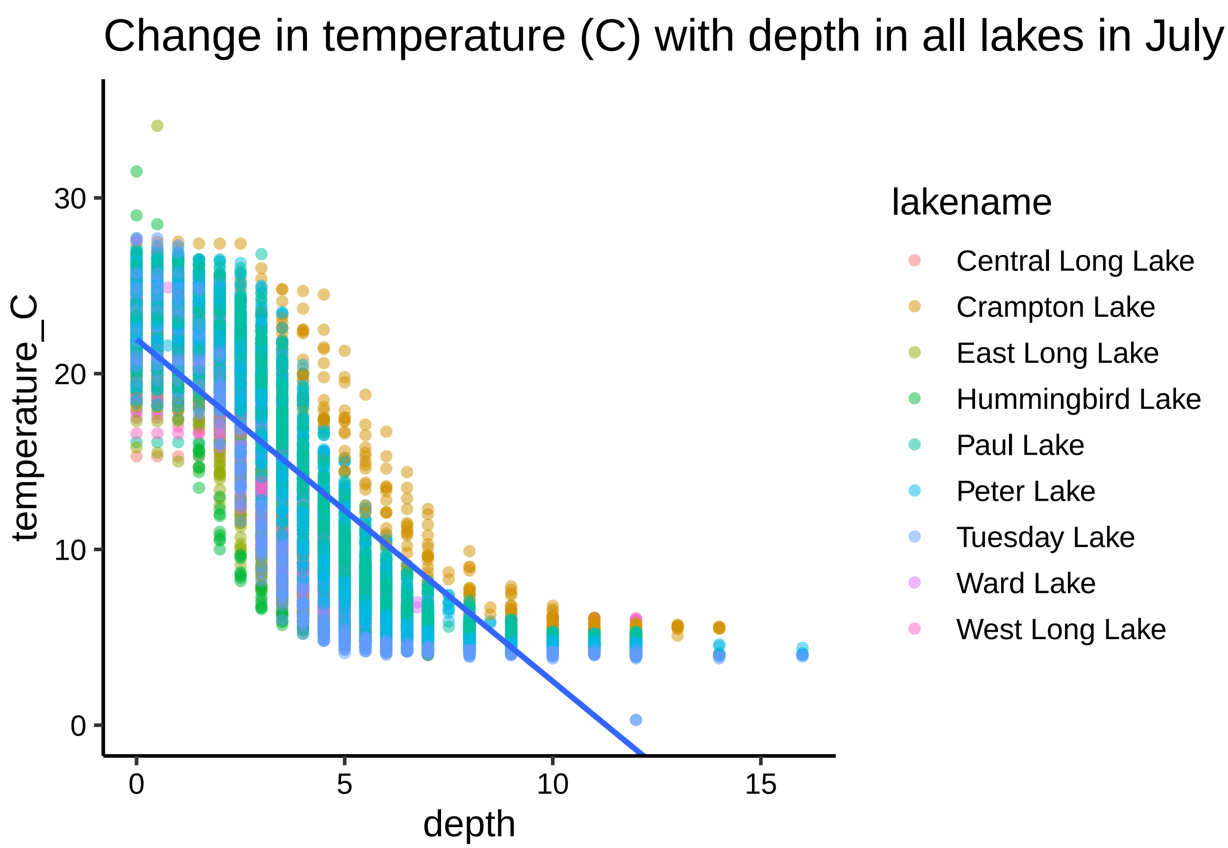
13. Is there a significant difference in mean temperature among the lakes? Report your findings.

    Answer: Yes, in both, the aov and glm tests, the overall p-value < 0.05 indicating that the null hypothesis should be rejected ie there is a significant difference in the mean temperature of the lakes. However, it does not indicate which pairs of lakes are statistically significant.

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.
ggplot(data_processed, aes(x=depth, y=temperature_C))+
  geom_point(aes(color = lakename), alpha = 0.5)+
  geom_smooth(method="lm", se=FALSE)+
  coord_cartesian(ylim=c(0,35))+
  ggtitle("Change in temperature (C) with depth in all lakes in July")
```

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



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

```
#15
print(TukeyHSD(tempvslakename_anova))
```

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

```r
print(HSD.test(tempvslakename_anova, "lakename", group=TRUE))
```

```
## $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:Peter Lake statistically has the same mean as Paul Lake and Ward Lake because the p-value > 0.05 and so we do not reject the null hypothesis that their means are equal (ie they are the same). This is supported by the HDF test. There is no lake that has a mean temperature that is statistically distinct from all 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:

the t-test can be used to compare the means of two samples.

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
CramptonWardLakes_July <- data_processed %>%
  filter(lakename %in% c("Crampton Lake", "Ward Lake"))

t.test(CramptonWardLakes_July$temperature_C ~ CramptonWardLakes_July$lakename)
```

```
##
##  Welch Two Sample t-test
##
## data:  CramptonWardLakes_July$temperature_C by CramptonWardLakes_July$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 r
## 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 p-value > 0.05 meaning the null hypothesis that their means are equal is not rejected ie they are statistically the same. This matches the result of the Tukey HSD test where their p-value is also > 0.05 and thus their difference in means is not significant.