IreneChang_A06_GLMs.Rmd

Irene Chang

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.

```
#1
getwd()
```

```
## [1] "/home/guest/EDA/EDA-Spring2023"
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.0
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.2.1     v stringr 1.5.0
## v readr   2.1.3     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 /home/guest/EDA/EDA-Spring2023
```

```
library(lubridate)
```

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

```
here()
```

```
## [1] "/home/guest/EDA/EDA-Spring2023"
```

```
NTL.Lake <- read.csv(here("Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv"), stringsAsFactors = TRUE)
NTL.Lake$sampledate <- mdy(NTL.Lake$sampledate)

#2
mytheme <-
  theme_classic() +
  theme(axis.text = element_text(color = "black", hjust = 0.5),
        plot.title = element_text(hjust = 0.5),
        legend.position = "bottom")
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 the depth across all lakes. Ha: The mean lake temperature recorded during July does change with the depth across all lakes.
4. Wrangle your NTL-LTER dataset with a pipe function so that the records meet the following criteria:

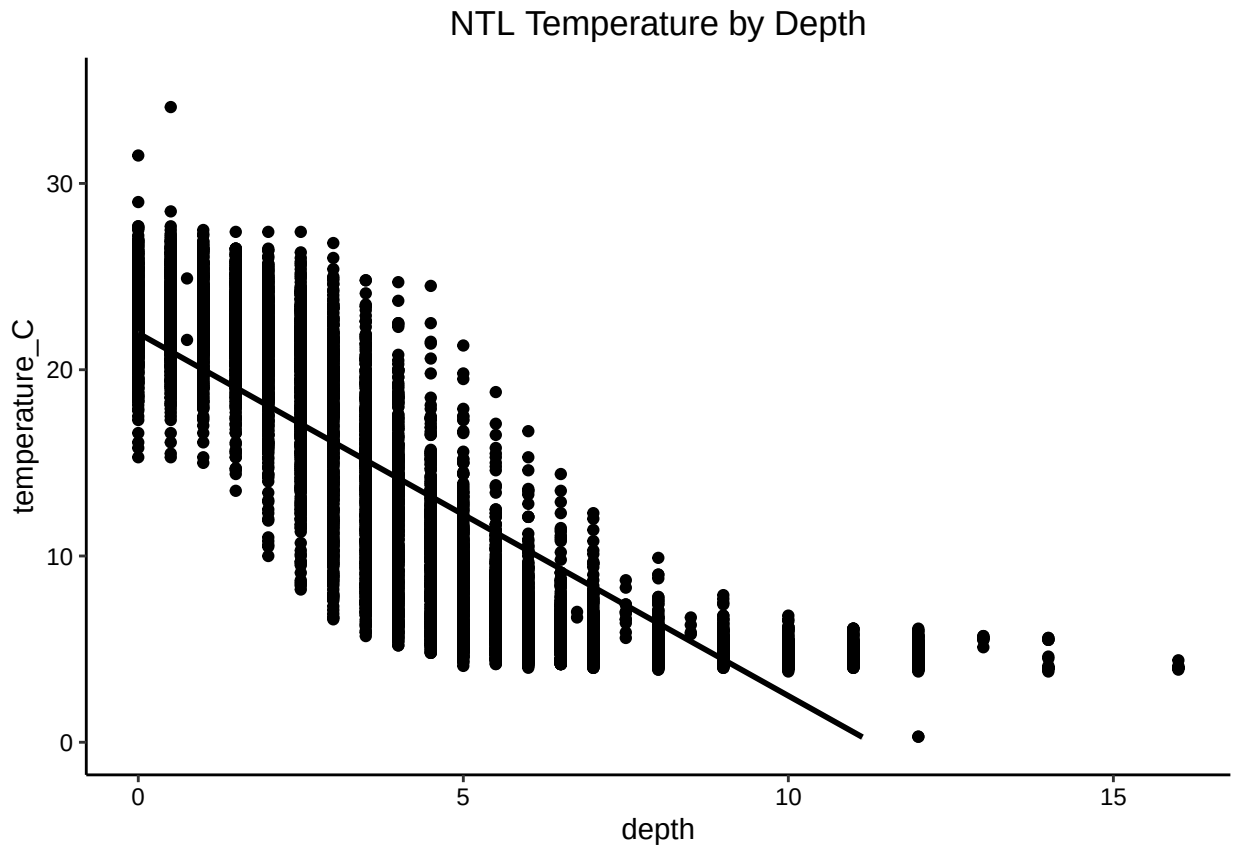
- Only dates in July.
- Only the columns: `lakename`, `year4`, `daynum`, `depth`, `temperature_C`
- Only complete cases (i.e., remove NAs)

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.LTER.month <- mutate(NTL.Lake, month = month(sampledate))

NTL.LTER.July <- NTL.LTER.month %>%
```

```
  filter(month == 7)

NTL.LTER.subset <- NTL.LTER.July %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  na.omit(lakename, year4, daynum, depth, temperature_C)

#5

NTL.tempbydepth <-
  ggplot(NTL.LTER.subset, aes(x=depth, y=temperature_C)) +
  ylim(0, 35) +
  geom_point() +
  geom_smooth(method = 'lm', color = "black") +
  labs(
    title = "NTL Temperature by Depth")
print(NTL.tempbydepth)
```

```
## `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?

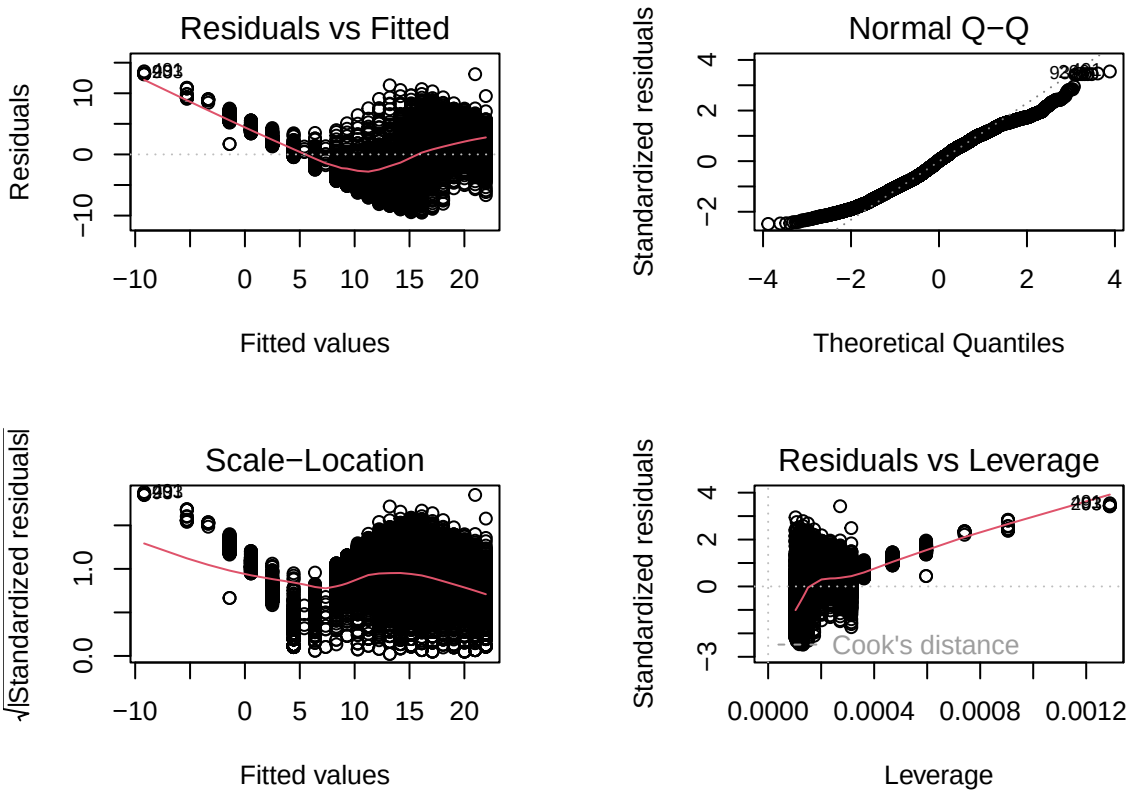
Answer: The plot suggests that the higher the depth (deeper), the lower the temperature of the lake is in celcius. The distribution of points suggests that depth vs temperature has a strong negative correlation. The distribution of points suggests that the relationship between depth vs temperature is relatively linear but the data also suggests that the best goodness of fit line may not be linear. When looking at the depth of 10 or above, the line does not accurately reflect the relationship between the two variables based on the line of best fit. This may suggest that a better goodness of fit between the two variables may be a curved line. There is also a high varianace outside of the line of best fit, which suggests that the correlation between the two variables may not be very high.

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

```
#7

temperature.regression <- lm(data = NTL.LTER.subset, temperature_C ~ depth)
summary(temperature.regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = NTL.LTER.subset)
##
## 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
```

```
par(mfrow = c(2,2), mar=c(4,4,4,4))
plot(temperature.regression)
```




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

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: 73.87% of the variability in temperature is explained by changes in depth. The degrees of freedom is 9,726. The result is statistically significant because the pvalue is 2.2e-16, which is less than 0.05. For a 1m change in depth, the temperature will decrease by 1.94621 degrees celcius.

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
NTL.LTER.AIC <- NTL.LTER.July %>%
  select(year4, daynum, depth, temperature_C) %>%
  na.omit(year4, daynum, depth, temperature_C)

tempAIC <- lm(data = NTL.LTER.AIC, temperature_C ~ depth + year4 + daynum)
step(tempAIC)
```

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

```r
#the set of explanatory variables that is best suited
#to predict temperature is one that includes all three of those variables.
#The AIC is 26066.

#10
temp.multipleregression <- lm(data = NTL.LTER.July, temperature_C ~ depth + year4 + daynum)
summary(temp.multipleregression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = NTL.LTER.July)
##
## 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
##   (1116 observations deleted due to missingness)
```

```
## 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 final set of explanatory variables that the AIC method suggests to predict temperature in our multiple regression are year4, daynum, and depth. 74.12% of the variability can be explained by these variables. It is an improvement over the model that only uses depth as an explanatory variable; in that case 73.87% of the variability was explained by just depth.

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
NTL.LTER.July.naomit <- NTL.LTER.July[!(is.na(NTL.LTER.July$temperature_C)), ]
summary(NTL.LTER.July.naomit$temperature_C)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.30    5.50   10.10   12.72   20.80   34.10
```

```
NTL.subset.lakename <- NTL.LTER.July.naomit %>%
  group_by(lakename, sampledate) %>%
  summarise(temperature_C = mean(temperature_C))
```

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

```
summary(NTL.subset.lakename)
```

```
##            lakename     sampledate         temperature_C
##  Paul Lake     :150   Min.   :1984-07-01   Min.   : 8.645
##  Peter Lake    :150   1st Qu.:1991-07-24   1st Qu.:11.190
##  Tuesday Lake  : 86   Median :1997-07-25   Median :12.700
##  West Long Lake: 52   Mean   :1999-02-16   Mean   :12.865
##  East Long Lake: 49   3rd Qu.:2006-07-03   3rd Qu.:14.150
##  Crampton Lake : 15   Max.   :2016-07-27   Max.   :21.022
##  (Other)       : 31
```

```
NTL.subset.lakename.anova <- aov(data = NTL.subset.lakename, temperature_C ~ lakename)
summary(NTL.subset.lakename.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
```

```
NTL.subset.lakename.anova2 <- lm(data = NTL.subset.lakename, temperature_C ~ lakename)
summary(NTL.subset.lakename.anova2)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = NTL.subset.lakename)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2754 -0.9386 -0.1271  0.7646  7.1552
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 17.6870     0.3864  45.769  < 2e-16 ***
## lakenameCrampton Lake       -2.3176     0.5373  -4.313 1.92e-05 ***
## lakenameEast Long Lake      -7.4019     0.4382 -16.892  < 2e-16 ***
## lakenameHummingbird Lake    -6.8655     0.6178 -11.113  < 2e-16 ***
## lakenamePaul Lake           -3.8099     0.4041  -9.429  < 2e-16 ***
## lakenamePeter Lake          -4.2466     0.4041 -10.509  < 2e-16 ***
## lakenameTuesday Lake        -6.4968     0.4167 -15.590  < 2e-16 ***
## lakenameWard Lake           -3.2574     0.6408  -5.083 5.18e-07 ***
## lakenameWest Long Lake      -6.1034     0.4354 -14.019  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.446 on 524 degrees of freedom
## Multiple R-squared:  0.5489, Adjusted R-squared:  0.542
## F-statistic: 79.69 on 8 and 524 DF,  p-value: < 2.2e-16
```

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

    Answer: The p-value for both the ANOVA and the linear regression are both below 0.05, therefore, it is statistically significant. Therefore, we can reject the null hypothesis and conclude that there is a significant difference in the mean temperatures between the lakes in the month of July.

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.

NTL.tempbydepth2 <-
  ggplot(NTL.Lake, aes(x=depth, y=temperature_C, color=lakename)) +
  ylim(0, 35) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = 'lm', se = FALSE, color = "black") +
```

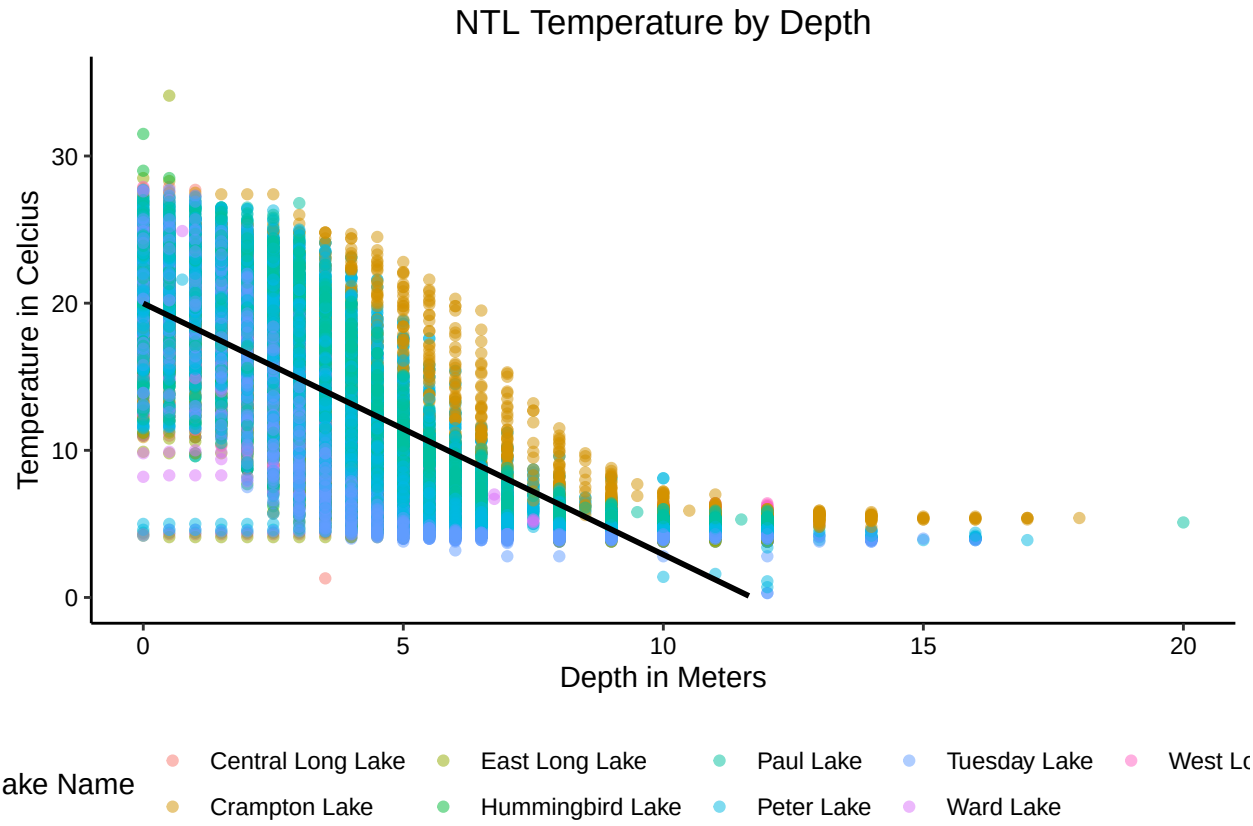
```
  labs(
    title = "NTL Temperature by Depth",
    x = "Depth in Meters",
    y = "Temperature in Celcius",
    color = "Lake Name")
print(NTL.tempbydepth2)
```

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

```
## Warning: Removed 3858 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 3858 rows containing missing values (`geom_point()`).
```

```
## Warning: Removed 33 rows containing missing values (`geom_smooth()`).
```




```
#used original dataset because does not specify only July months.
```

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

```
#15

TukeyHSD(NTL.subset.lakename.anova)
```

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

```
subset.lakename.groups <- HSD.test(NTL.subset.lakename.anova, "lakename", group = TRUE)
subset.lakename.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: Both Paul Lake and Ward Lake have the same mean temperature, statistically speaking, as Peter Lake. For both Paul Lake and Ward lake have the same mean temperature, statistically speaking, because their p-value is greater than 0.05, so we fail to reject the null hypothesis that the have the same mean temperature. Central Long Lake has a mean temperature that is statistically distinct from all the other lakes because the difference in means of Central Long Lake and all the other lakes has a p-value of less than 0.05. This means that we reject the null hypothesis that the difference in means is 0.

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 also use the two-sample T test because a two-sample T test tests the hypothesis that the two lakes would have the same mean temperature.

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?

```
NTL.LTER.specificlakes <- NTL.LTER.July %>%
  filter(lakename %in% c("Crampton Lake", "Ward Lake"))

Julytemp.twosample <- t.test(NTL.LTER.specificlakes$temperature_C ~ NTL.LTER.specificlakes$lakename)
Julytemp.twosample
```

```
##
##  Welch Two Sample t-test
##
## data:  NTL.LTER.specificlakes$temperature_C by NTL.LTER.specificlakes$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: We do not reject the null hypothesis. We cannot conclude that the difference of the mean temperatures for Crampton Lake and Ward Lake is not equal to 0 because the p-value is greater than 0.05 at 0.2649. This supports the results in part 16 because the interaction between the two lakes in the Tukey's HSD test showed the p-value of the difference in means to be greater than 0.05. Therefore, in both tests, we fail to reject the null hypothesis that the difference in means is 0.