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

Vicky Fong

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
install.packages("agricolae")
```

```
## Installing package into '/home/guest/R/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library(tidyverse); library(lubridate); library(here); library(agricolae); library(dplyr)
```

```
## -- 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 errors
## here() starts at /home/guest/EDE_Fall2024
```

```r
getwd()
```

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

```r
ntl <- read.csv(
  here('Data','Raw','NTL-LTER_Lake_ChemistryPhysics_Raw.csv')
  )

ntl$year4 <- as.factor(ntl$year4)
ntl$sampledate <- mdy(ntl$sampledate)

#2
my_theme <-  theme(
    plot.title = element_text(face = "bold", size = 12),
    panel.background = element_rect(fill = "white", colour = NA),
    panel.border = element_rect(fill = NA, colour="grey50"),
    panel.grid.major = element_line(colour = "black", size = 0.01),
    panel.grid.minor =  element_line(colour = "black", size = 0.01),
    axis.text = element_text(size = 10),
    axis.ticks = element_blank(),
  )
```

```
## Warning: The ‘size‘ argument of ‘element_line()‘ is deprecated as of ggplot2 3.4.0.
## i Please use the ‘linewidth‘ argument instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

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 lake temperature recorded during July will not change with depth across all lakes. Ha: Mean lake temperature recorded during July will decrease as depth increaeses 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.

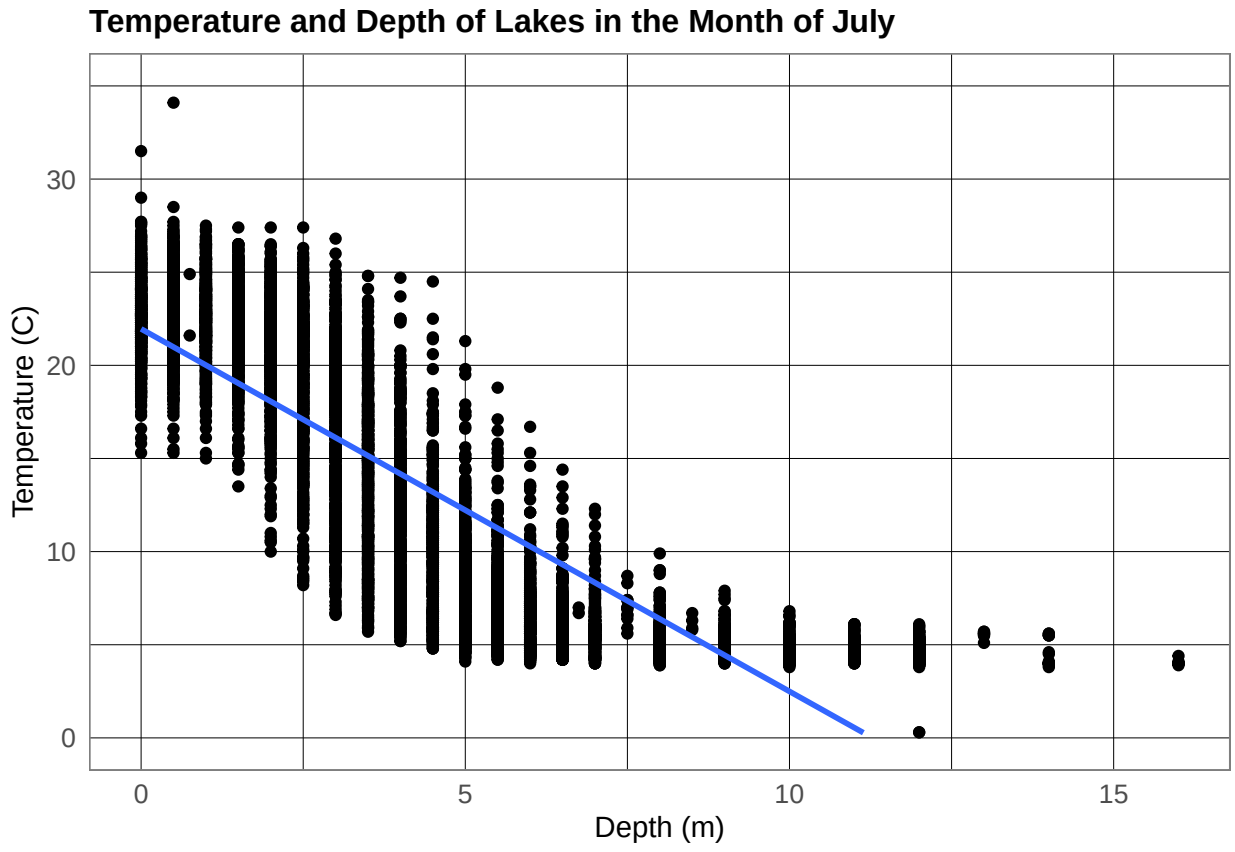
```r
#4
ntl <- ntl %>%
  mutate(month = month(sampledate)) %>%
  filter(month == 7) %>%
```

```
  select(lakename,year4,daynum,depth,temperature_C) %>%
  na.omit()

#5
ntl %>%
  ggplot(aes(x = depth, y = temperature_C)) +
  geom_point() +
  geom_smooth(method=lm, se=FALSE) +
  labs(x="Depth (m)", y="Temperature (C)",
       title="Temperature and Depth of Lakes in the Month of July") +
  ylim(0,35) +
  my_theme
```

```
## `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 supports our alternative hypothesis that temperature would decrease with increasing depth. The inverse relationship between temperature and depth is most evident from 0 to 5 meters depth, and the strong negative slope suggessts that the greatest temperature change

occurs in the first 5 meters of the lake depth. The range of temperatures between 5 to 10 meters is smaller, suggesting less temperture changes occur in this depth range. From 10 to 15 meters, most of the temperature points are around 5 degrees C.

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

```
#7
ntl.regression <- lm(data = ntl, temperature_C ~ depth)
summary(ntl.regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = ntl)
##
## 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 R-squared value of 0.7387 means that 73.9% of the variabilitty is explained by changes in depth. This finding is based ono 9726 degrees of freedonm. Since the P-value is less than 0.05, it suggests a statistically significant relationship between temperature and depth of lakes in the month of July. The coefficient of the slope is -1.94621, which means tempeature is predicted to decrease by 1.95 degrees C for every 1 meter increase in depth.

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.AIC <- lm(data = ntl, temperature_C ~ depth + year4 + daynum)
step(ntl.AIC) #include all variables
```

```
## Start:  AIC=25378.53
## temperature_C ~ depth + year4 + daynum
##
##          Df Sum of Sq    RSS   AIC
## <none>                131187 25379
## - daynum  1      1359 132546 25477
## - year4  32     10601 141788 26070
## - depth   1    402237 533424 39022
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = ntl)
##
## Coefficients:
## (Intercept)        depth    year41985    year41986    year41987    year41988
##    14.69898     -1.94808     -0.59193     -0.81263     -0.29746     -0.14561
##   year41989    year41990    year41991    year41992    year41993    year41994
##    -1.25525     -1.04234     -1.13797     -2.30326     -1.95840     -0.86579
##   year41995    year41996    year41997    year41998    year41999    year42000
##    -1.45019     -2.76107     -2.20265     -0.93812     -2.05150     -2.82785
##   year42001    year42002    year42003    year42004    year42005    year42006
##    -1.14400     -1.41667     -0.15619     -0.80519      2.17728      0.13083
##   year42007    year42008    year42009    year42010    year42011    year42012
##     0.28831     -0.77024     -0.75285      0.70631      0.19033      0.42349
##   year42013    year42014    year42015    year42016       daynum
##    -2.27570     -2.64737     -1.13700     -1.31197      0.04208
```

```r
#10
summary(ntl.AIC)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = ntl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3850 -2.8904  0.0839  2.8207 13.0231
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 14.698977   0.866986   16.954  < 2e-16 ***
## depth       -1.948080   0.011300 -172.395  < 2e-16 ***
## year41985   -0.591931   0.336591   -1.759 0.078676 .
## year41986   -0.812629   0.343411   -2.366 0.017984 *
## year41987   -0.297462   0.296431   -1.003 0.315655
## year41988   -0.145608   0.349239   -0.417 0.676740
## year41989   -1.255251   0.322432   -3.893 9.96e-05 ***
## year41990   -1.042336   0.310370   -3.358 0.000787 ***
```

```
## year41991   -1.137972   0.295029   -3.857 0.000115 ***
## year41992   -2.303257   0.302801   -7.607 3.08e-14 ***
## year41993   -1.958401   0.303264   -6.458 1.11e-10 ***
## year41994   -0.865789   0.304028   -2.848 0.004413 **
## year41995   -1.450188   0.299174   -4.847 1.27e-06 ***
## year41996   -2.761072   0.303201   -9.106  < 2e-16 ***
## year41997   -2.202650   0.295123   -7.464 9.15e-14 ***
## year41998   -0.938119   0.306513   -3.061 0.002215 **
## year41999   -2.051501   0.293775   -6.983 3.07e-12 ***
## year42000   -2.827850   0.428461   -6.600 4.33e-11 ***
## year42001   -1.143999   0.343728   -3.328 0.000877 ***
## year42002   -1.416669   0.315858   -4.485 7.37e-06 ***
## year42003   -0.156188   0.312560   -0.500 0.617294
## year42004   -0.805190   0.313257   -2.570 0.010173 *
## year42005    2.177278   0.335610    6.488 9.15e-11 ***
## year42006    0.130829   0.313873    0.417 0.676819
## year42007    0.288314   0.352909    0.817 0.413968
## year42008   -0.770236   0.363399   -2.120 0.034071 *
## year42009   -0.752852   0.376108   -2.002 0.045346 *
## year42010    0.706309   0.344106    2.053 0.040139 *
## year42011    0.190333   0.383602    0.496 0.619784
## year42012    0.423486   0.323569    1.309 0.190635
## year42013   -2.275702   0.327460   -6.950 3.90e-12 ***
## year42014   -2.647366   0.330704   -8.005 1.33e-15 ***
## year42015   -1.136996   0.332227   -3.422 0.000623 ***
## year42016   -1.311974   0.341709   -3.839 0.000124 ***
## daynum       0.042082   0.004199   10.021  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.679 on 9693 degrees of freedom
## Multiple R-squared:  0.7604, Adjusted R-squared:  0.7595
## F-statistic: 904.6 on 34 and 9693 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 recommended including all available variables (depth, year and day number) to predict temperature in the multiple regression as the step function only provided one option. Multiple R-squared is 0.7604, suggesting this model explains 76.0% of the observed vaariance. This is a slight improvement from the model with depth as the only variable at 73.9%.

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$lakename <- as.factor(ntl$lakename)
ntl.anova <- aov(data = ntl, temperature_C ~ lakename)
summary(ntl.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
```

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

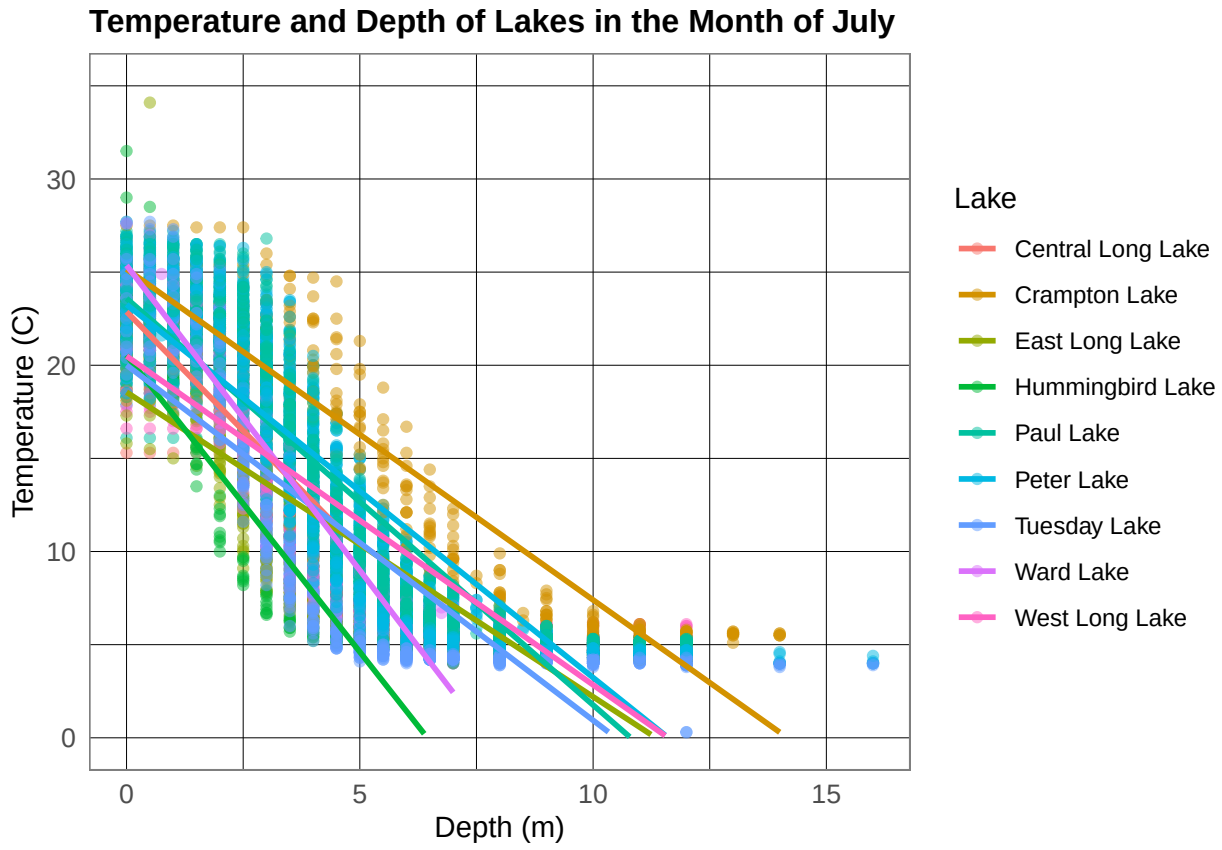
    Answer: Yes, the p-value of the anova is less than 0.05, suggesting that there is a significant difference in mean temperature among the lakes.

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.
ntl %>%
  ggplot(aes(x = depth, y = temperature_C, color = lakename)) +
  geom_point(alpha=0.5) +
  geom_smooth(method=lm, se=FALSE) +
  labs(x="Depth (m)", y="Temperature (C)", color="Lake",
       title="Temperature and Depth of Lakes in the Month of July") +
  ylim(0,35) +
  my_theme
```

```
## `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
TukeyHSD(ntl.anova)
```

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

```
ntl.groups <- HSD.test(ntl.anova, "lakename", group = TRUE)
ntl.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: The Tukey's HSD test found that Ward Lake and Peter Lake had the same mean temperature, statistically speaking, as Peter Lake. There were no lakes that had a statistically distinct mean temperature.

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 can use two-sample t-test to see if Peter Lake and Paul Lake have statistically distinct mean temperatures.

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 <- ntl %>%
  filter(lakename == c("Crampton Lake", "Ward Lake"))
two.t <- t.test(crampton.ward$temperature_C ~ crampton.ward$lakename)
two.t
```

```
##
##  Welch Two Sample t-test
##
## data:  crampton.ward$temperature_C by crampton.ward$lakename
## t = 0.98673, df = 95.77, p-value = 0.3263
## alternative hypothesis: true difference in means between group Crampton Lake and group Ward Lake is n
## 95 percent confidence interval:
##  -1.130614  3.365610
```

```
## sample estimates:
## mean in group Crampton Lake     mean in group Ward Lake
##                     15.37107                     14.25357
```

Answer: The two-sample t-test had a p-value of 0.3263, which is greater than 0.05, thus we are unable to reject our null hypothesis. This suggests that there is no statistical difference between the July temperatures in Crampton Lake and Ward Lake. The mean temperatures of the lakes are not equal - 15.4 for Crampton Lake and 14.3 for Ward Lake - but not different enough to be statistically significant. This matches the results from the Tukey HSD as Crampton Lake and Ward Lake both belonged in group "b", which means that mean temperatures are statistically similar enough for them to be grouped together for analysis.