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

Emma Wellbaum

OVERVIEW

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

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Fay_A06_GLMs.Rmd") prior to submission.

The completed exercise is due on Tuesday, March 2 at 1:00 pm.

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
# Check working directory
getwd()
```

```
## [1] "C:/Users/emmaw/Documents/ENV872/Environmental_Data_Analytics_2021"
```

```
# Load needed packages
library(tidyverse)
library(agricolae)
library(lubridate)
# Import raw NTL-LTER data for chemistry/physics
NTL.phys.data <- read.csv("./Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv")
# Format sampledate as date
NTL.phys.data$sampledate <- as.Date(NTL.phys.data$sampledate, format = "%m/%d/%y")

#2
# Create theme
mytheme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
# Set theme as default
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 lake temperature recorded during July does not change with depth across all lakes. Ha: Mean lake temperature recorded during July changes with depth across all lakes.

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

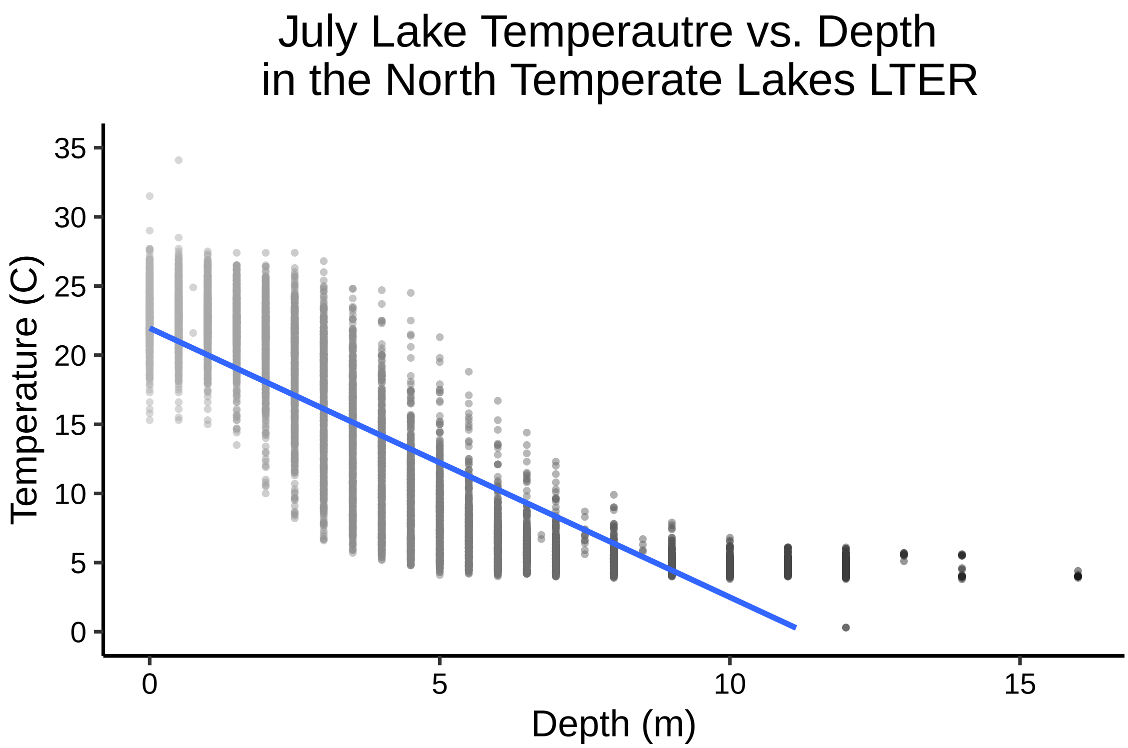
- Only dates in July.
- Only the columns: `lakename`, `year4`, `daynum`, `depth`, `temperature_C`
- Only complete cases (i.e., remove NAs)

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.

```r
#4
NTL.phys.July <-
  NTL.phys.data %>%
  # Create a month column based on 'sampledate'
  mutate(Month = month(sampledate)) %>%
  # Filter to dates in July only
  filter(Month == 7) %>%
  # Subset the data to the specified columns
  select(lakename, year4, daynum, depth, temperature_C) %>%
  # Remove incomplete records
  na.omit()

#5
tempbydepth.plot <-
  ggplot(data = NTL.phys.July, aes(x= depth, y=temperature_C, color = depth)) +
  geom_point(size = 0.7, alpha = 0.5) +
  # Limit temperatures to values from 0 to 35 degrees and set breaks
  scale_y_continuous(limits = c(0, 35), breaks = c(0,5,10,15,20,25,30,35)) +
  # Set colors
  scale_color_gradient(low="grey70", high="grey10") +
  # Add a smoothed line showing a linear model
  geom_smooth(method = "lm") +
  # Add axis labels and title
  labs(x="Depth (m)", y="Temperature (C)",
       title="July Lake Temperautre vs. Depth \n in the North Temperate Lakes LTER") +
  # Center title
  theme(plot.title = element_text(hjust=0.5), legend.position = "none")
print(tempbydepth.plot)
```

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



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 suggests that temperature and depth are negatively correlated. As depth increases from the surface of the water temperature also decreases. While a linear model does not appear to be a poor fit here, the distribution of points suggests that this trend is not truly linear. The symmetry of the model falls appart as depth increases until there are no points below the trend line at all. This makes sense given our research question. If the relationship between temperature and depth were truly linear then the water at the bottom of lakes freeze once it reached 0 degrees Celcius. One alternative option would be to take the log of temperture and see if resulting trend line is a better fit for the data.

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

```r
#7
# Simple linear regression
tempbydepth.regression <- lm(data = NTL.phys.July, temperature_C ~ depth)
summary(tempbydepth.regression)
```

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

```
# Correlation test
cor.test(NTL.phys.July$temperature_C, NTL.phys.July$depth)
```

```
##
##  Pearson's product-moment correlation
##
## data:  NTL.phys.July$temperature_C and NTL.phys.July$depth
## 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
```

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: There is a statistically significant negative correlation between temperature and depth (p-value <0.001). The model, which is based on 9726 degrees of freedom, predicts that temperautre will change -1.94621 degrees Celcius for every 1m change in depth. According to the model, changes in depth explain 73.87% of the variability in temperature.

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
# Run a stepwise AIC to determine what set of explanatory variables is best
# suited to predict temperature

TempAIC <- lm(data = NTL.phys.July, temperature_C ~ year4 + daynum + depth)
step(TempAIC)
```

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

```
#10
# Run a multiple regression on the recommended set of variables
temp.regression <- lm(data = NTL.phys.July, temperature_C ~ year4 + daynum + depth)
summary(temp.regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL.phys.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
## 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 suggest that we use year4, daynum, and depth (all three explanatory variables) to predict temperature in our multiple regression. This multiple regression model explains roughly 74.12% of the variance in temperature based on 9724 degrees of freedom. This model is an improvement over using depth as the only explanatory variable, as indicated in part by a larger R-squared value and lower residual standard error.

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

```
# Linear model
NTL.phys.July.anova.lm <- lm(data = NTL.phys.July, temperature_C ~ lakename)
summary(NTL.phys.July.anova.lm)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = NTL.phys.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.

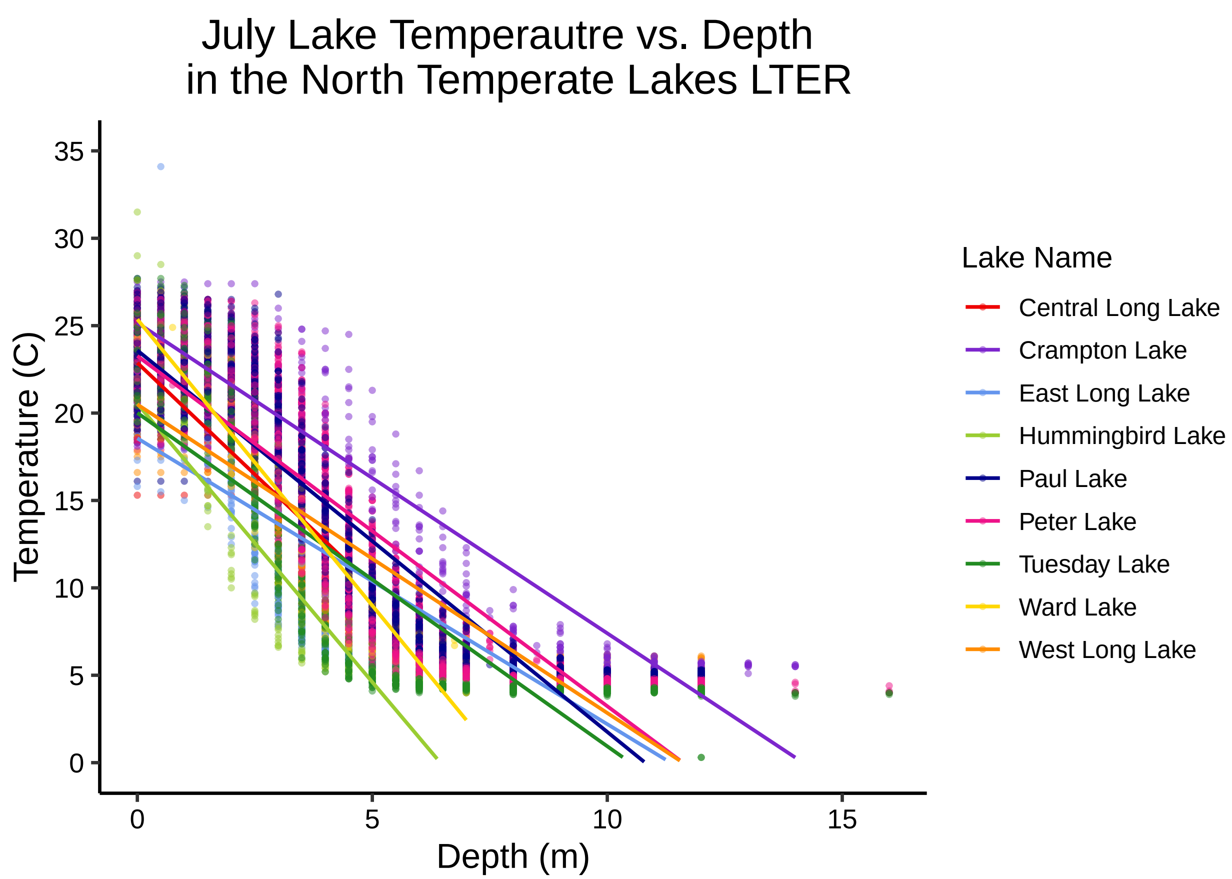
    Answer: Yes. Based on the ANOVA model, we can tell that there is a significant difference in mean temperature across all of the NTL-LTER lakes collectively (p-value <0.001). The ANOVA test expressed as a linear model indicates that there is a siginificant difference between the mean temperatures of each individual lake in the NTL-LTER (p-values <0.01 or 0.001). Thus, we should reject the null hypothesis and conclude that not all lakes in the NTL-LTER have the same mean temperature, on average, during 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.
lakename.plot <-
  ggplot(data = NTL.phys.July, aes(x= depth, y=temperature_C, color=lakename)) +
  # Make points 50% transparent
  geom_point(size = 0.7, alpha = 0.5) +
  # Limit temperatures to values from 0 to 35 degrees and set breaks
  scale_y_continuous(limits = c(0, 35), breaks = c(0,5,10,15,20,25,30,35)) +
  # Set colors
```

```
  scale_color_manual(values = c("red2", "purple3", "cornflowerblue",
                                "yellowgreen", "darkblue", "deeppink2",
                                "forestgreen", "gold1", "darkorange")) +
  # Add a smoothed linear regression for each lake
  geom_smooth(method = "lm", se = FALSE, size=0.7) +
  # Add axis and legend labels and a plot title
  labs(x="Depth (m)",
       y="Temperature (C)",
       color="Lake Name",
       title="July Lake Temperautre vs. Depth \n in the North Temperate Lakes LTER") +
  # Center title and format legend
  theme(plot.title = element_text(hjust=0.5),
        legend.position = "right",
        legend.title = element_text(size=12),
        legend.text = element_text(size=10),
        legend.margin = margin(0, 0, 0, 0, "pt"))
print(lakename.plot)
```

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



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

```
#15
TukeyHSD(NTL.phys.July.anova)
```

```
##   Tukey multiple comparisons of means
```

```
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = NTL.phys.July)
##
## $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
lake.groups <- HSD.test(NTL.phys.July.anova, "lakename", group = TRUE)
lake.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 Min  Max    Q25   Q50    Q75
## Central Long Lake      17.66641 4.196292  128 8.9 26.8 14.400 18.40 21.000
## Crampton Lake          15.35189 7.244773  318 5.0 27.5  7.525 16.90 22.300
## East Long Lake         10.26767 6.766804  968 4.2 34.1  4.975  6.50 15.925
## Hummingbird Lake       10.77328 7.017845  116 4.0 31.5  5.200  7.00 15.625
## Paul Lake              13.81426 7.296928 2660 4.7 27.7  6.500 12.40 21.400
## Peter Lake             13.31626 7.669758 2872 4.0 27.0  5.600 11.40 21.500
## Tuesday Lake           11.06923 7.698687 1524 0.3 27.7  4.400  6.80 19.400
## Ward Lake              14.45862 7.409079  116 5.7 27.6  7.200 12.55 23.200
## West Long Lake         11.57865 6.980789 1026 4.0 25.7  5.400  8.00 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: Ward Lake and Paul Lake have the same mean temperature ("c") as Peter Lake, statistically speaking. There is no lake that has a mean temperature that is statistically distinct from all other lakes. The lakes all share at least one group letter with another 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 run a two-sample t-test to explore whether Peter Lake and Paul Lake have distinct mean temperatures. A two-sample t-test is used to test whether two samples have the same mean.