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

Jon Nichols

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 Monday, February 28 at 7:00 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

getwd()
```

```
## [1] "C:/Users/nicho/OneDrive/Documents/ENV872/Environmental_Data_Analytics_2022"
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(agricolae)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

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

Lake$sampledate <- as.Date(Lake$sampledate , format = "%m/%d/%y")

#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: Temperature = a + b1(Depth) + e H0: b1 = 0 Ha: b1 != 0
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

Lake.subset <-
  Lake %>%
  mutate(month = month(sampledate)) %>%
  filter(month == 07) %>%
  select(lakename:daynum, depth:temperature_C) %>%
  na.omit()

summary(Lake.subset$temperature_C)
```

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

```
summary(Lake.subset$depth)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.000   4.500   4.745   7.000  16.000
```

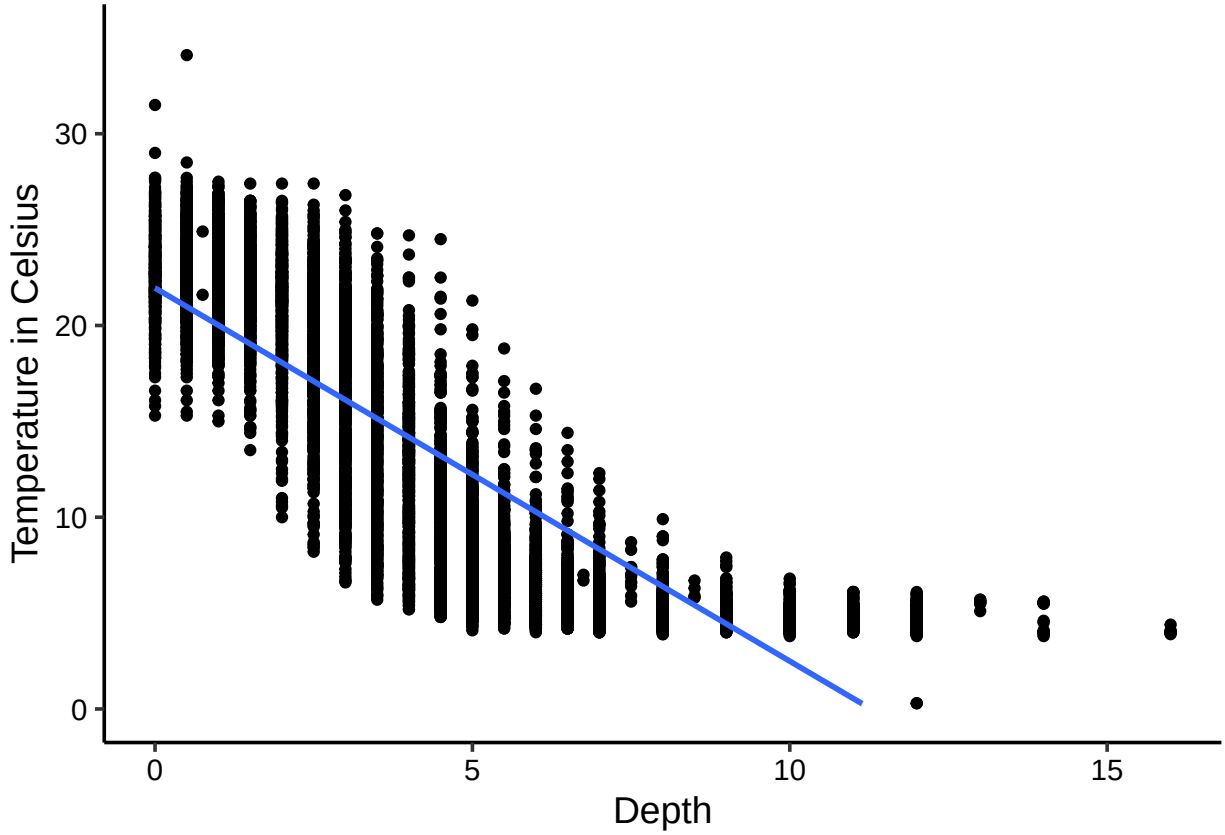
```
#5

tempvsdepth <-
  ggplot(Lake.subset, aes(x = depth, y = temperature_C)) +
  geom_point() +
  geom_smooth(method = lm)+
```

```
  ylab("Temperature in Celsius")+
  xlab("Depth")+
  ylim(0, 35)
print(tempvsdepth)
```

```
## `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: As depth increases, temperature decreases. The distribution of points appear to be linearly correlated up to a depth of 10m, at which point the temperature hits a limit of approximately 5C even as depth continues to increase.

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

```
#7

lake.regression <- lm(data = Lake.subset, temperature_C ~ depth)
summary(lake.regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = Lake.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
```

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: Depth explains 73.87% of the variability in temperature. There are 9,726 degrees of freedom on which this finding is based. The p-value for depth is less than 0.05, indicating that we can reject the null hypothesis and our coefficient for the relationship between depth and temperature is statistically different than zero. The p-value for the entire regression is also lower than the 0.05 confidence level, indicating the regression is meaningful and can help explain the relationship between temperature and 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.

```
#9

NTLAIC <- lm(data = Lake.subset, temperature_C ~ depth + year4 + daynum)
step(NTLAIC)
```

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

```
#10

AICmodel <- lm(data = Lake.subset, temperature_C ~ depth + year4 + daynum)
summary(AICmodel)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = Lake.subset)
##
## 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
## 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: Depth, year, and day number were included in the final set of explanatory variables. The model explains 74.12% of the observed variance in temperature. This is only a slight improvement over using depth as the sole explanatory variable, and the increase in the R2 is likely due to adding two additional independent variables rather than accounting for additional variance.

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

Lake.Totals <- Lake.subset %>%
  group_by(lakename, year4,daynum)%>%
  summarise(meantemp = mean(temperature_C))
```

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

```
Lake.Totals.anova <- aov(data = Lake.Totals, meantemp ~ lakename)
summary(Lake.Totals.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
```

```
Lake.Totals.anova2 <- lm(data = Lake.Totals, meantemp ~ lakename)
summary(Lake.Totals.anova2)
```

```
##
## Call:
## lm(formula = meantemp ~ lakename, data = Lake.Totals)
##
## 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: Yes, both anova tests have p-levels less than 0.05, thus we reject the null hypothesis.

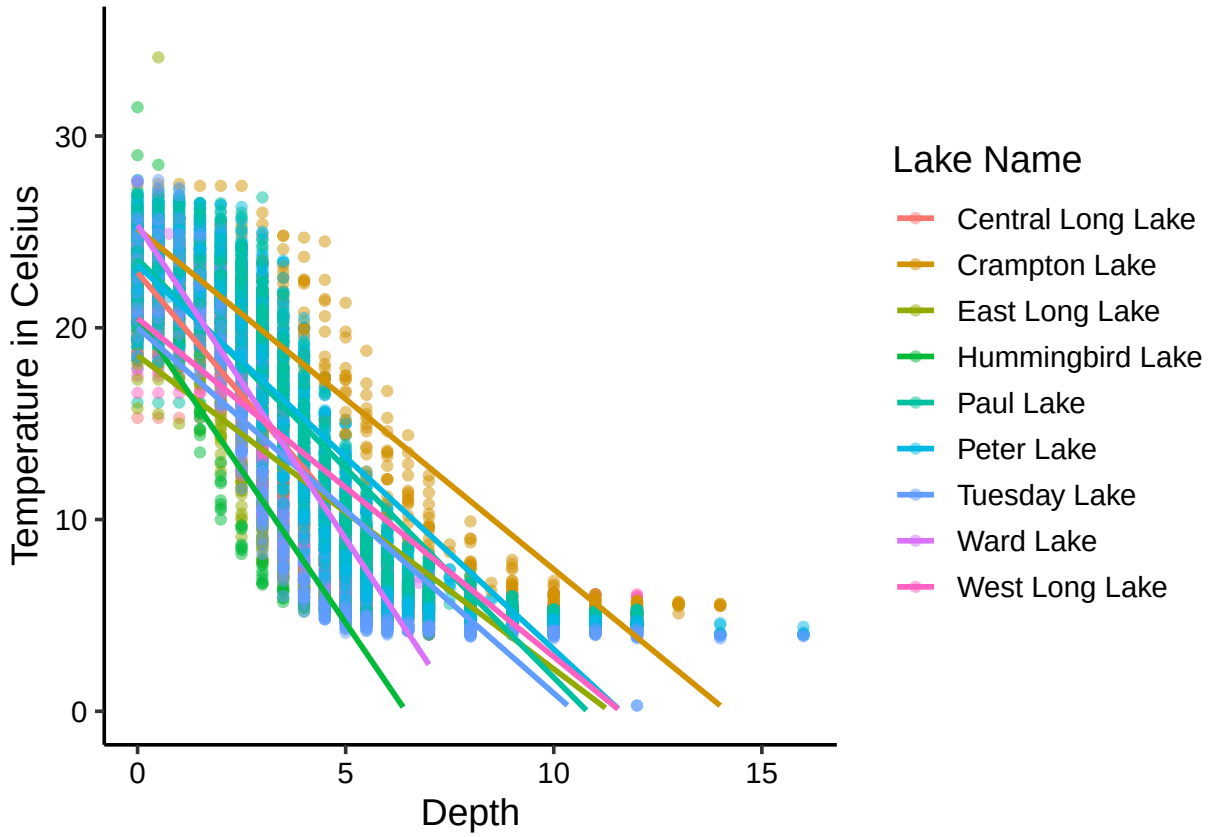
14. Create a graph that depicts temperature by depth, with a separate color for each lake. Add a geom_smooth (method = "lm", se = FALSE) for each lake. Make your points 50 % transparent. Adjust your y axis limits to go from 0 to 35 degrees. Clean up your graph to make it pretty.

```
#14.

tempvsdepth2 <-
  ggplot(Lake.subset, aes(x = depth, y = temperature_C, color=lakename)) +
  geom_point(alpha = .5) +
  geom_smooth(method = lm, se=FALSE)+
  labs(color = "Lake Name")+
  ylab("Temperature in Celsius")+
  xlab("Depth")+
  ylim(0, 35)
print(tempvsdepth2)
```

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

```
## Warning: Removed 73 rows containing missing values (geom_smooth).
```



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

```
#15

TukeyHSD(Lake.Totals.anova)
```

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

```
Lake.totals.groups <- HSD.test(Lake.Totals.anova, "lakename", group = TRUE)
Lake.totals.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
##                   meantemp       std   r       Min      Max      Q25      Q50
## Central Long Lake 17.68698 1.9058275  14 14.544444 21.02222 16.66389 17.61889
## Crampton Lake     15.36943 1.5138376  15 12.909091 18.10476 14.47045 15.11818
## East Long Lake    10.28505 1.0519538  49  8.645000 12.70000  9.49000 10.08500
## Hummingbird Lake  10.82151 0.7962931   9  9.146667 11.88462 10.45385 10.97692
## Paul Lake         13.87706 1.2358897 150 10.900000 17.80625 12.97778 13.85000
## Peter Lake        13.44040 1.8099936 150 10.165000 19.54000 12.14708 13.24500
## Tuesday Lake      11.19023 1.4616471  86  8.935000 18.34545 10.43281 10.91111
## Ward Lake         14.42956 1.5572801   8 12.360000 16.10714 13.31141 14.54282
## West Long Lake    11.58357 0.9263405  52  9.745000 13.49500 10.99250 11.58000
##                       Q75
## Central Long Lake 18.71389
## Crampton Lake     16.25000
## East Long Lake    10.98500
## Hummingbird Lake  11.16154
## Paul Lake         14.57500
## Peter Lake        14.55319
## Tuesday Lake      11.70937
## Ward Lake         15.77156
## West Long Lake    12.25500
```

```
##
## $comparison
## NULL
##
## $groups
##                    meantemp 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: Ward Lake and Paul 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 can do a two tailed t-test. Null hypothesis would be mean(peter) - mean(paul) = 0. Alternative hypothesis would be that the mean is not equal to 0.