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

Natalie von Turkovich

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] "/Users/natalievonturkovich/Documents/DUKE/Courses/Spring 22/ENV_872_EDA/Environmental_Data_Analy
```

```
library(tidyverse)
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     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
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
NTL_LTR_ChemPhys<-read.csv("/Users/natalievonturkovich/Documents/DUKE/Courses/Spring 22/ENV_872_EDA/Envi

NTL_LTR_ChemPhys$sampledate <- mdy(NTL_LTR_ChemPhys$sampledate)

#2
mytheme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "right") #alternative: legend.position + legend.justification

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: There is no change in mean lake temperature in July across depths Ha: There is is a difference in means of lake temperature in July across depths

null is always there is no change alt is always there is a change

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.

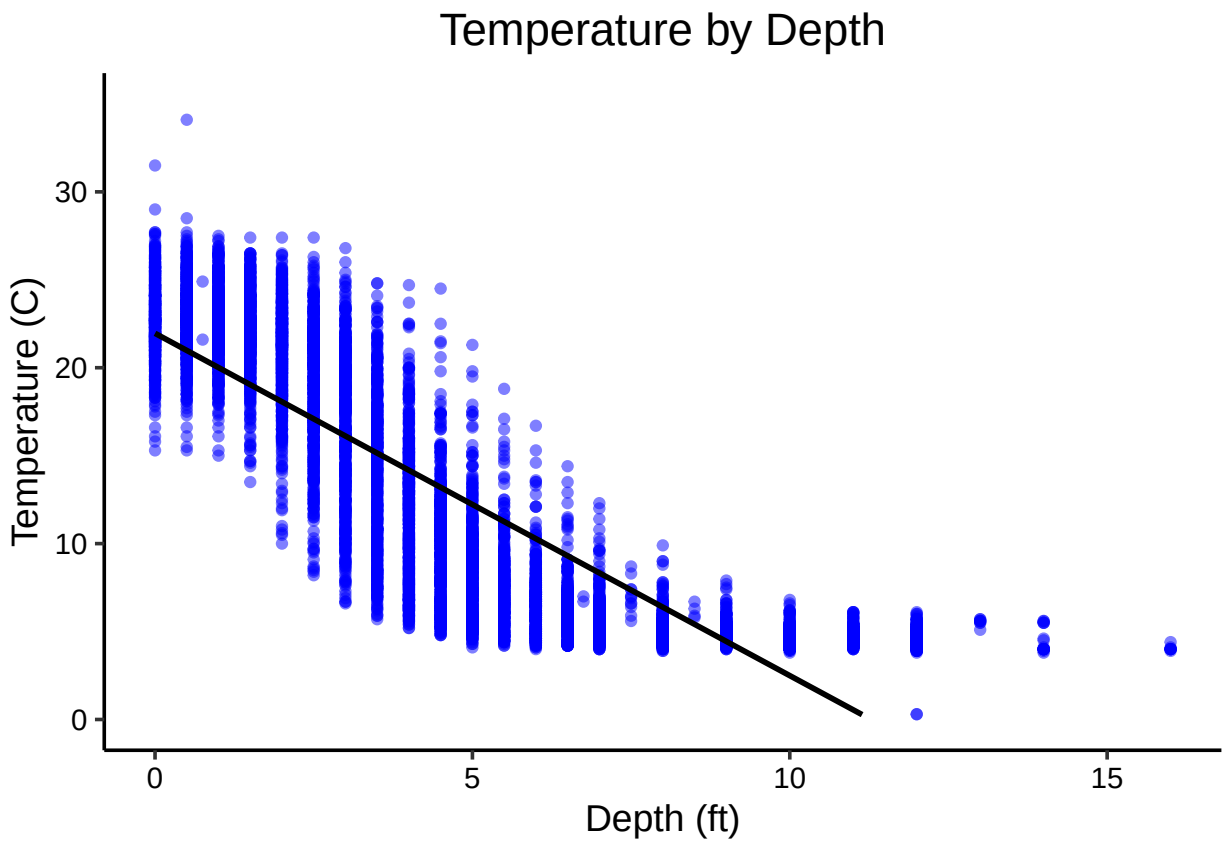
```r
#4

NTL_LTR_ChemPhys.processed <-NTL_LTR_ChemPhys %>%
          mutate(Month = month(sampledate)) %>%
          filter(Month == 7) %>%
          select(c(lakename:daynum, depth, temperature_C)) %>%
          drop_na()


#5
ggplot(NTL_LTR_ChemPhys.processed, aes(x=depth, y = temperature_C))+
  geom_point(color = "blue", alpha=.5)+
  geom_smooth(method="lm", se=F, color = 'black')+ #se=F gets rid of SE grey shaded area
  ylim(0,35)+
  labs(x= "Depth (ft)", y= "Temperature (C)", title = "Temperature by Depth")+
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `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?

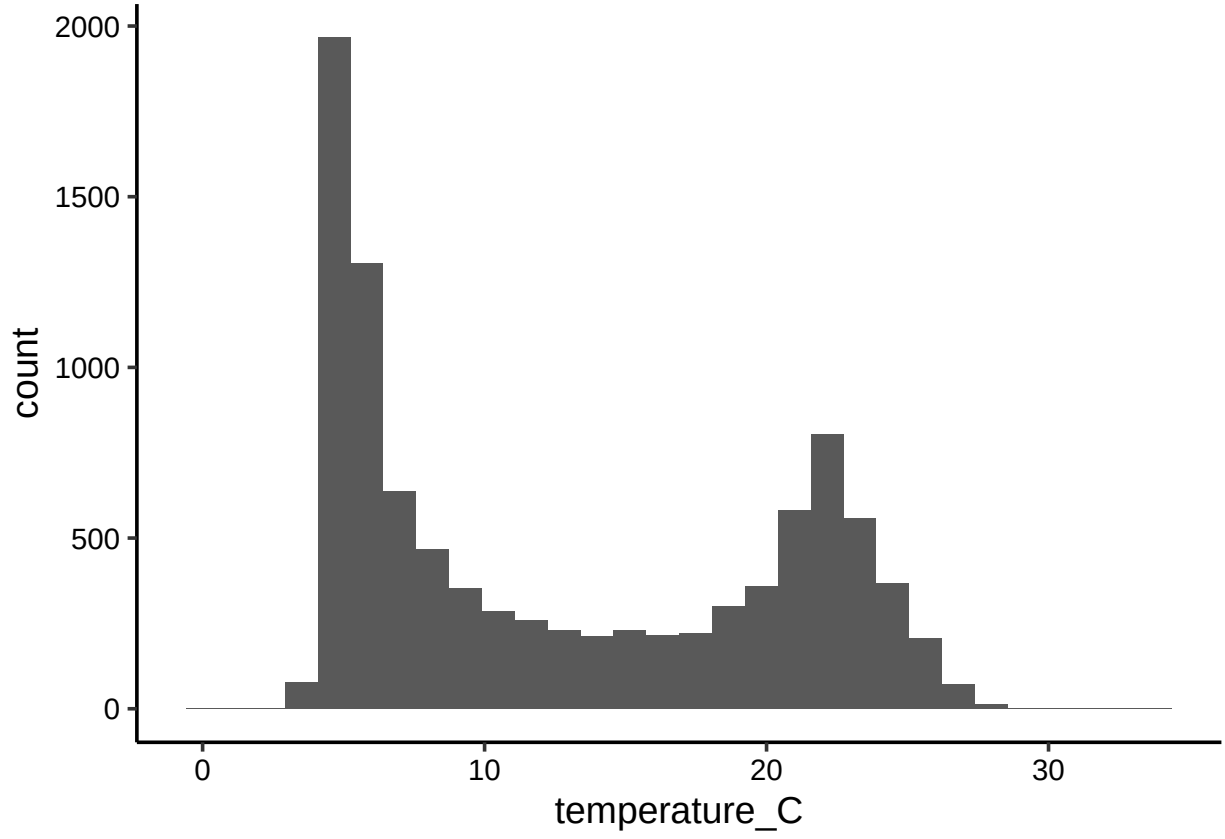
Answer: The figure suggests that as depth increases temperature decreases. Once you get to ~9ft in depth, temperature stays pretty consistent around ~6-7 degrees C. The distribution of points suggest that there is a negative linear trend.

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

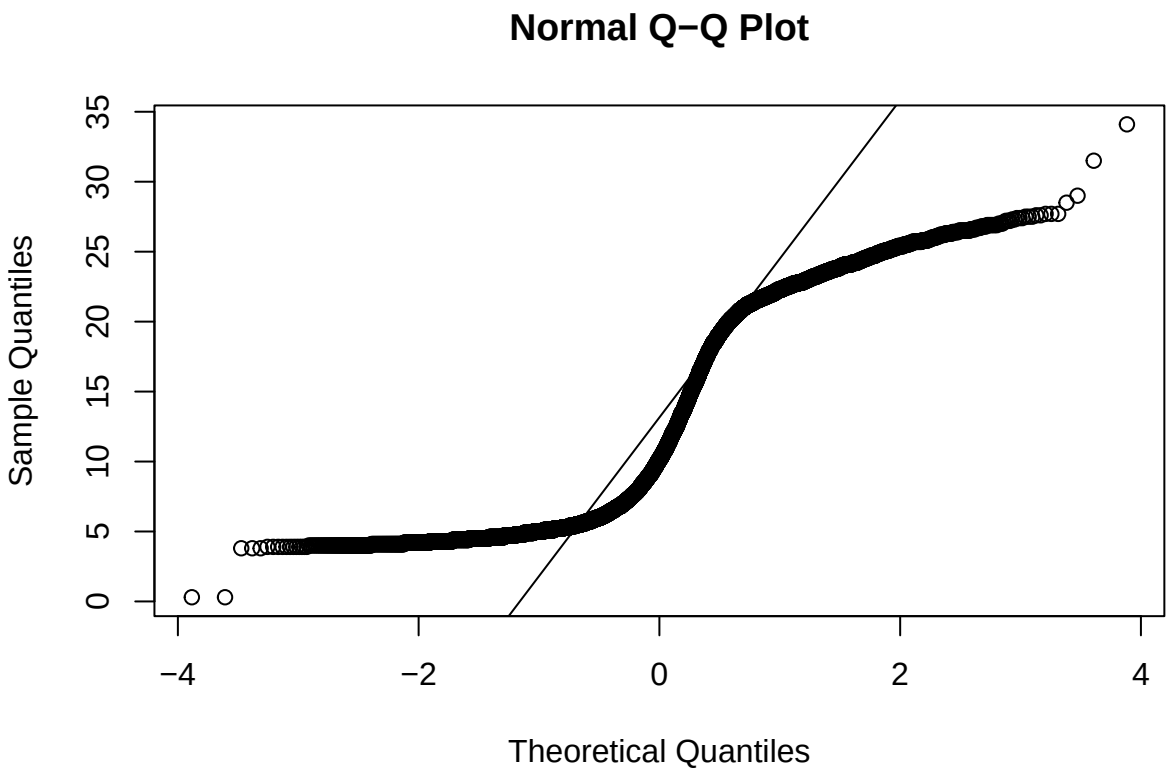
```
#7
#Did not run shapiro - ran the tests below instead
#NTL_LTR_ChemPhys.subsample<-sample(NTL_LTR_ChemPhys.processed, 50)
#shapiro.test(NTL_LTR_ChemPhys.processed$temperature_C)
#p value is low - that means data is not normal


ggplot(NTL_LTR_ChemPhys.processed, aes(x = temperature_C)) +
  geom_histogram() #not evenly distributed
```

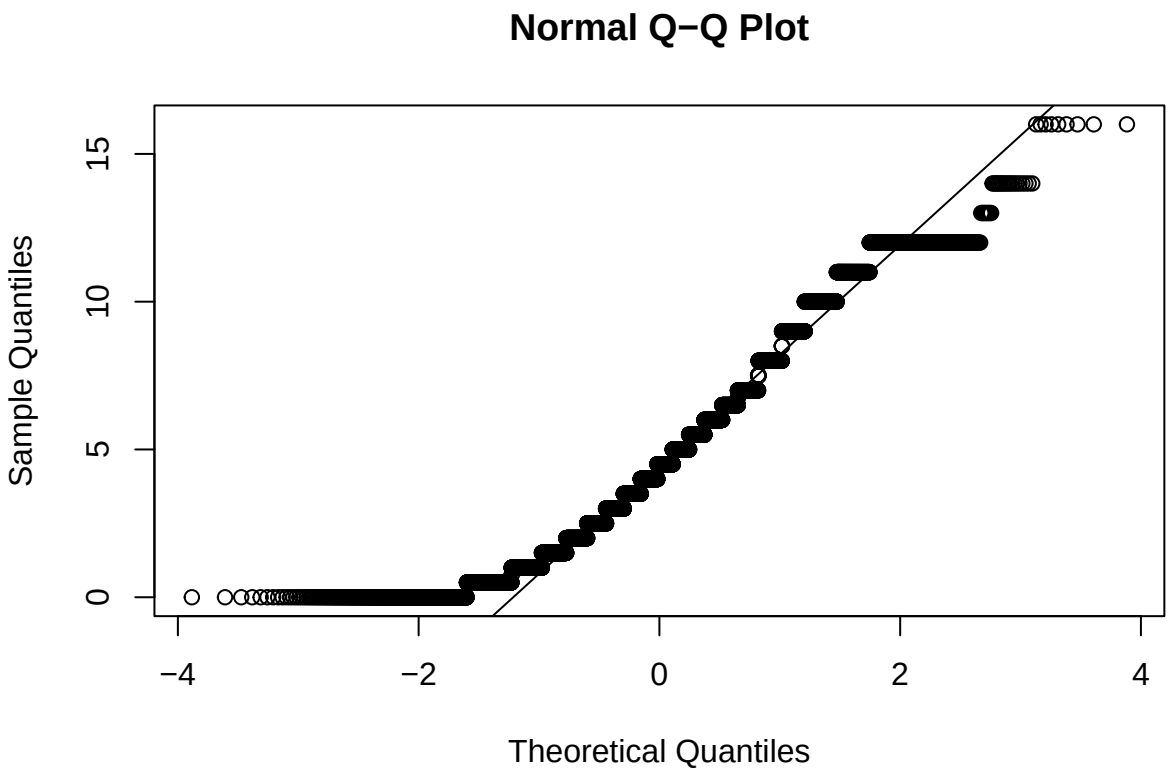
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
qqnorm(NTL_LTR_ChemPhys.processed$temperature_C); qqline(NTL_LTR_ChemPhys.processed$temperature_C) #sho
```




```
qqnorm(NTL_LTR_ChemPhys.processed$depth); qqline(NTL_LTR_ChemPhys.processed$depth) #depth better but no
```




```
NTL_LTR_ChemPhys.regression <- lm(data = NTL_LTR_ChemPhys.processed, temperature_C ~ depth)
summary(NTL_LTR_ChemPhys.regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = NTL_LTR_ChemPhys.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 results show a high R squared value, indicating that 73% for temperature is explained by depth, with 9726 degrees of freedom. The low p value <.05 is statistically significant, we can reject the null hypothesis that there is no change in mean lake temperature in July across depths. We can interpret the depth coefficient where with every 1 meter increase in depth results in a 1.9 degree C decrease in temperature.

y=21.95-1.94X y changes by x

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
NTL_LTR_ChemPhys.AIC <- lm(data = NTL_LTR_ChemPhys.processed, temperature_C ~ depth + year4 +
              daynum)

#Choose a model by AIC in a Stepwise Algorithm
step(NTL_LTR_ChemPhys.AIC)
```

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

```
#the AIC shows that the lowest AIC score is when all three of these variables
#are included in the regression

#10
NTL_LTR_ChemPhys.regression <- lm(data = NTL_LTR_ChemPhys.processed, temperature_C ~ depth + year4 +
              daynum)
summary(NTL_LTR_ChemPhys.regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = NTL_LTR_ChemPhys.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
## 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: The AIC method revealed that depth, year and daynum should all be included in the regression. The R squared value tells us that 74% of the observed variance is explained by the model. This is an improvement from the 73% of of the observed variance explained from our previous model with just depth as an independent variable.

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
# Wrangle the data

NTL_LTR_ChemPhys.totals<-NTL_LTR_ChemPhys.processed %>%
  group_by(lakename, temperature_C) %>%
  summarise(tempsum = sum(temperature_C))
```

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

```
summary(NTL_LTR_ChemPhys.totals)
```

```
##            lakename   temperature_C       tempsum
##  Peter Lake     :231   Min.   : 0.300   Min.   :  0.60
##  Paul Lake      :228   1st Qu.: 9.425   1st Qu.: 22.93
##  Tuesday Lake   :222   Median :15.150   Median : 51.25
##  West Long Lake:191   Mean   :15.207   Mean   : 86.78
##  East Long Lake:187   3rd Qu.:20.900   3rd Qu.:106.47
##  Crampton Lake :136   Max.   :34.100   Max.   :794.50
##  (Other)       :231
```

```
#check all value of plotID
summary(NTL_LTR_ChemPhys.totals$lakename)
```

```
## Central Long Lake     Crampton Lake    East Long Lake  Hummingbird Lake
##                75               136               187                71
##         Paul Lake        Peter Lake      Tuesday Lake         Ward Lake
##               228               231               222                85
##    West Long Lake
##               191
```

```
# Format ANOVA as aov
NTL_LTR_ChemPhys.anova <- aov(data = NTL_LTR_ChemPhys.totals, temperature_C ~ lakename)
summary(NTL_LTR_ChemPhys.anova)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## lakename       8    935  116.92   2.747 0.0052 **
## Residuals   1417  60313   42.56
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#results: reject null hypothesis i.e. difference between a pair of group means is stat. significant
#P<.05,  null = mean is same across all sites. But doesn't tell us which means are not the same.

# Format ANOVA as lm
NTL_LTR_ChemPhys.anova2 <- lm(data = NTL_LTR_ChemPhys.totals, temperature_C ~ lakename)
summary(NTL_LTR_ChemPhys.anova2)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = NTL_LTR_ChemPhys.totals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.8550  -5.7529  -0.1518   5.6458  19.5503
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 16.8360     0.7533  22.348  < 2e-16 ***
## lakenameCrampton Lake       -1.4375     0.9383  -1.532 0.125765
## lakenameEast Long Lake      -2.2863     0.8917  -2.564 0.010452 *
## lakenameHummingbird Lake    -4.1614     1.0803  -3.852 0.000122 ***
## lakenamePaul Lake           -0.8609     0.8684  -0.991 0.321698
## lakenamePeter Lake          -1.3360     0.8671  -1.541 0.123576
```

```
## lakenameTuesday Lake      -1.6810     0.8714  -1.929 0.053900 .
## lakenameWard Lake         -1.5160     1.0336  -1.467 0.142664
## lakenameWest Long Lake    -2.0821     0.8890  -2.342 0.019320 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.524 on 1417 degrees of freedom
## Multiple R-squared:  0.01527,    Adjusted R-squared:  0.009712
## F-statistic: 2.747 on 8 and 1417 DF,  p-value: 0.005197
```

```
#different way of running test
```

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

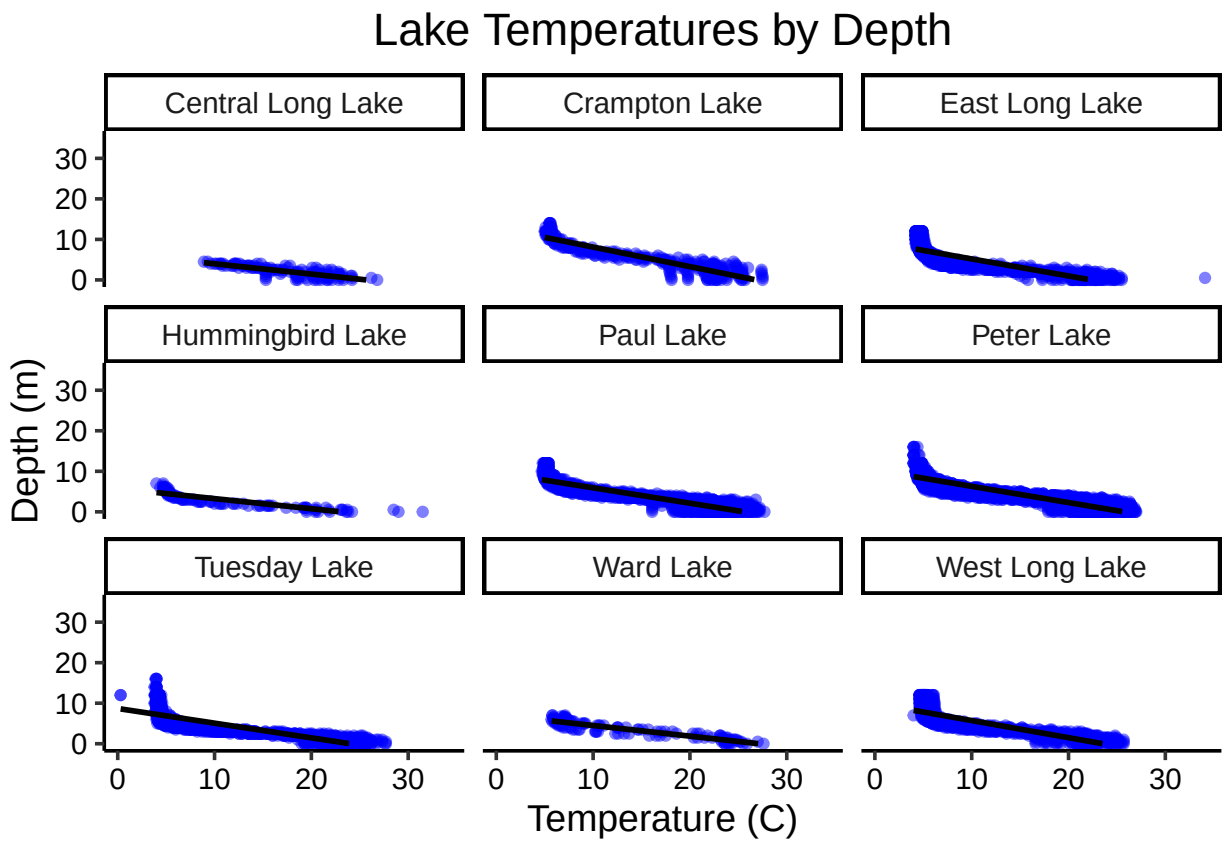
    Answer: P <.05, we can reject the null hypothesis the that mean temperature is the same across all lakes. This means that there is a significant difference in mean temperature among the lakes. A mulitple R squared value of .015 indicates that than only 1.5% of mean temperature is explained by which lake it is.

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.
# Graph the results
ggplot(NTL_LTR_ChemPhys.processed, aes(x= temperature_C, y = depth))+
  geom_point(alpha = .5, color = "blue")+
  ylim(0,35)+
  geom_smooth(method = "lm", se = FALSE, color = "black")+
  facet_wrap(vars(lakename))+
  #scale_color_viridis_c()+
  labs(x= "Temperature (C)", y = "Depth (m)", title = "Lake Temperatures by Depth")+
  theme(plot.title = element_text(hjust = 0.5))
```

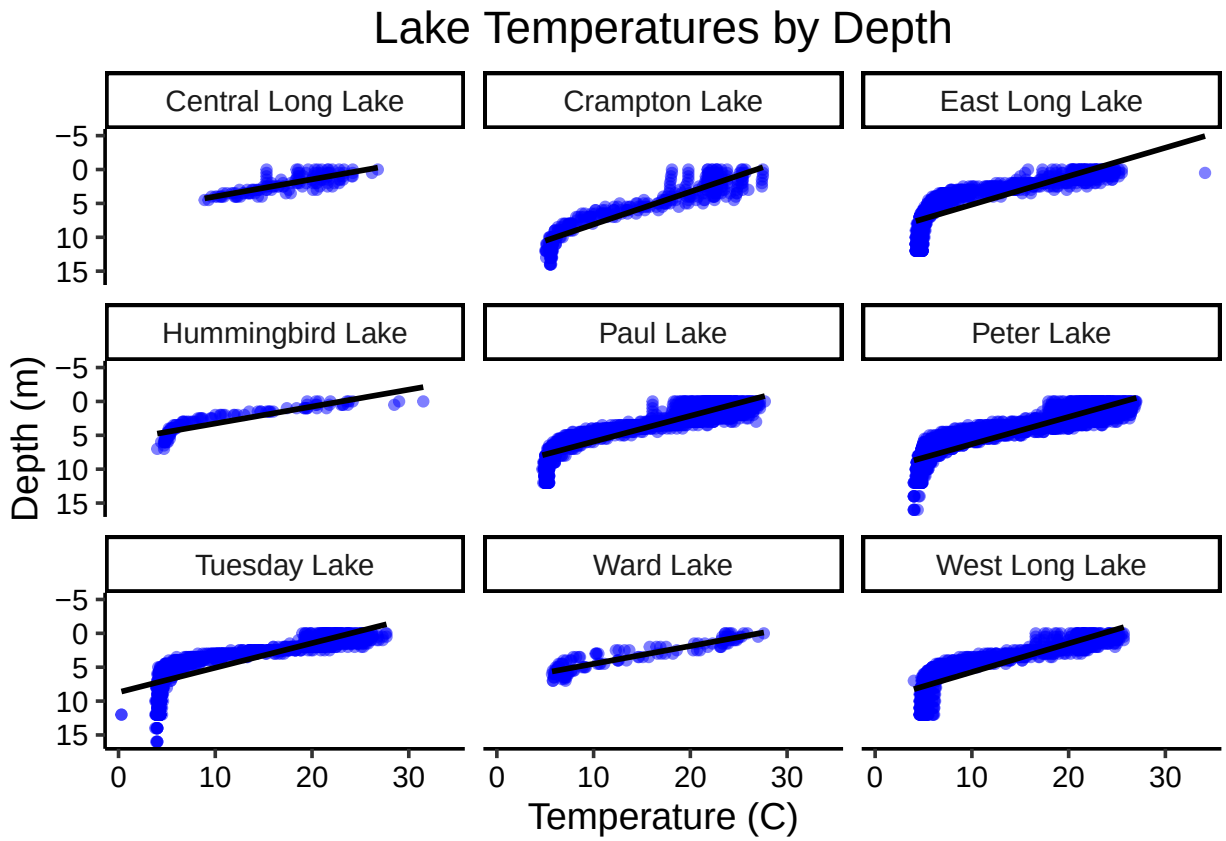
```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 99 rows containing missing values (geom_smooth).
```




```
ggplot(NTL_LTR_ChemPhys.processed, aes(x= temperature_C, y = depth))+
  geom_point(alpha = .5, color = "blue")+
  ylim(0,35)+
  geom_smooth(method = "lm", se = FALSE, color = "black")+
  scale_y_reverse()+
  facet_wrap(vars(lakename))+
  #scale_color_viridis_c()+
  labs(x= "Temperature (C)", y = "Depth (m)", title = "Lake Temperatures by Depth")+
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Scale for 'y' is already present. Adding another scale for 'y', which will
## replace the existing scale.
## `geom_smooth()` using formula 'y ~ x'
```



I reversed the x and y axis for this scatter plot. I find that it is a more natural way to represent lake temperature over depth as it depicts the cross section of a lake with the shallowest depth at the top of the y axis. I understand that temperature is the dependent variable and that the dependent variable typically goes on the y axis, but for this representation I put temperature on the x axis.

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

```
#15
# TukeyHSD() computes Tukey Honest Significant Differences
#identifies which means are same/different
TukeyHSD(NTL_LTR_ChemPhys.anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = NTL_LTR_ChemPhys.totals)
##
## $lakename
##                                          diff         lwr        upr     p adj
## Crampton Lake-Central Long Lake    -1.43747059 -4.35249623  1.4775551 0.8404834
## East Long Lake-Central Long Lake   -2.28626738 -5.05639928  0.4838645 0.2029304
## Hummingbird Lake-Central Long Lake -4.16135211 -7.51732048 -0.8053837 0.0038890
## Paul Lake-Central Long Lake        -0.86091228 -3.55880388  1.8369793 0.9867098
## Peter Lake-Central Long Lake       -1.33600000 -4.02955178  1.3575518 0.8360463
## Tuesday Lake-Central Long Lake     -1.68104505 -4.38794586  1.0258558 0.5934071
## Ward Lake-Central Long Lake        -1.51600000 -4.72685826  1.6948583 0.8708107
```

```
## West Long Lake-Central Long Lake   -2.08207330 -4.84388929  0.6797427 0.3175779
## East Long Lake-Crampton Lake       -0.84879679 -3.13288168  1.4352881 0.9655108
## Hummingbird Lake-Crampton Lake     -2.72388152 -5.69136139  0.2435983 0.1017310
## Paul Lake-Crampton Lake             0.57655831 -1.61935430  2.7724709 0.9964296
## Peter Lake-Crampton Lake            0.10147059 -2.08910793  2.2920491 1.0000000
## Tuesday Lake-Crampton Lake         -0.24357446 -2.45054639  1.9633975 0.9999946
## Ward Lake-Crampton Lake            -0.07852941 -2.88085413  2.7237953 1.0000000
## West Long Lake-Crampton Lake       -0.64460271 -2.91859494  1.6293895 0.9939526
## Hummingbird Lake-East Long Lake    -1.87508473 -4.70036224  0.9501928 0.5002106
## Paul Lake-East Long Lake            1.42535510 -0.57421696  3.4249272 0.3966606
## Peter Lake-East Long Lake           0.95026738 -1.04344536  2.9439801 0.8646350
## Tuesday Lake-East Long Lake         0.60522233 -1.40648871  2.6169334 0.9909770
## Ward Lake-East Long Lake            0.77026738 -1.88101157  3.4215463 0.9928576
## West Long Lake-East Long Lake       0.20419408 -1.88082316  2.2892113 0.9999979
## Paul Lake-Hummingbird Lake          3.30043983  0.54595596  6.0549237 0.0063640
## Peter Lake-Hummingbird Lake         2.82535211  0.07511876  5.5755855 0.0387168
## Tuesday Lake-Hummingbird Lake       2.48030707 -0.28300151  5.2436157 0.1196089
## Ward Lake-Hummingbird Lake          2.64535211 -0.61320171  5.9039059 0.2221523
## West Long Lake-Hummingbird Lake     2.07927881 -0.73784558  4.8964032 0.3470030
## Peter Lake-Paul Lake               -0.47508772 -2.36714451  1.4169691 0.9973880
## Tuesday Lake-Paul Lake             -0.82013276 -2.73114551  1.0908800 0.9213511
## Ward Lake-Paul Lake                -0.65508772 -3.23079508  1.9206196 0.9971418
## West Long Lake-Paul Lake           -1.22116102 -3.20919655  0.7668745 0.6081311
## Tuesday Lake-Peter Lake            -0.34504505 -2.24992609  1.5598360 0.9997585
## Ward Lake-Peter Lake               -0.18000000 -2.75116132  2.3911613 0.9999999
## West Long Lake-Peter Lake          -0.74607330 -2.72821541  1.2360688 0.9627677
## Ward Lake-Tuesday Lake              0.16504505 -2.42009738  2.7501875 0.9999999
## West Long Lake-Tuesday Lake        -0.40102825 -2.40127279  1.5992163 0.9994861
## West Long Lake-Ward Lake           -0.56607330 -3.20866235  2.0765158 0.9991646
```

```
# Extract groupings for pairwise relationships
NTL_LTR_ChemPhys.anova.groups <- HSD.test(NTL_LTR_ChemPhys.anova, "lakename", group = TRUE)
NTL_LTR_ChemPhys.anova.groups
```

```
## $statistics
##    MSerror   Df     Mean       CV
##   42.56406 1417 15.20689 42.90237
##
## $parameters
##    test   name.t ntr StudentizedRange alpha
##   Tukey lakename   9         4.393339  0.05
##
## $means
##                   temperature_C      std   r Min  Max    Q25   Q50    Q75
## Central Long Lake      16.83600 4.560750  75 8.9 26.8 12.750 17.00 20.550
## Crampton Lake          15.39853 6.462289 136 5.0 27.5  9.675 15.40 21.325
## East Long Lake         14.54973 6.426829 187 4.2 34.1  8.850 14.30 20.350
## Hummingbird Lake       12.67465 7.174991  71 4.0 31.5  6.550 10.50 18.800
## Paul Lake              15.97509 6.568314 228 4.7 27.7 10.275 15.95 21.625
## Peter Lake             15.50000 6.682814 231 4.0 27.0  9.750 15.50 21.250
## Tuesday Lake           15.15495 6.826118 222 0.3 27.7  9.225 15.25 20.975
## Ward Lake              15.32000 7.161135  85 5.7 27.6  8.200 14.70 23.200
## West Long Lake         14.75393 6.147107 191 4.0 25.7  9.350 14.90 20.150
##
```

```
## $comparison
## NULL
##
## $groups
##                    temperature_C groups
## Central Long Lake      16.83600      a
## Paul Lake              15.97509      a
## Peter Lake             15.50000      a
## Crampton Lake          15.39853     ab
## Ward Lake              15.32000     ab
## Tuesday Lake           15.15495     ab
## West Long Lake         14.75393     ab
## East Long Lake         14.54973     ab
## Hummingbird Lake       12.67465      b
##
## attr(,"class")
## [1] "group"
```

```
#indicates which groups have the same mean. Treatments with same letter are not sig. different
```

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:From the findings, there are three groups with the same mean temperatures. Those groups are: Group 1: Central Long Lake, Paul Lake, Peter Lake Group 2: Crampton Lake, Ward Lake, Tuesday Lake, West Long Lake, East Long Lake Group 3: Hummingbird Lake Peter Lake is in group 1 and has the same mean temperature as the other lakes in that group. Hummingbrid lake has a mean temperature that is statistically different from all other lakes at 12.67 degrees C.

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: If we were just looking at Peter and Paul Lake we could use a two-sample t test to test the if the mean of two lakes is distinct.