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

Sashoy Milton

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.

## Load packages

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- 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
```

```
## Check my working directory

getwd()
```

```
## [1] "C:/Users/sasho/Desktop/Environ Data Analytics/Env872 Workspace/EDA-Fall2022_SM/Assignments"
```

```
setwd("C:/Users/sasho/Desktop/Environ Data Analytics/Env872 Workspace/EDA-Fall2022_SM")

## Load data

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

## Set date variable

class(lake.chemistry$sampledate)  #Check date variable
```

```
## [1] "factor"
```

```
lake.chemistry$sampledate <- mdy(lake.chemistry$sampledate)
class(lake.chemistry$sampledate)  #Check date variable
```

```
## [1] "Date"
```

```
## Explore data

summary(lake.chemistry)
```

```
##      lakeid                lakename         year4          daynum
##  R      :11288   Peter Lake    :11288   Min.   :1984   Min.   : 55.0
##  L      :10325   Paul Lake     :10325   1st Qu.:1991   1st Qu.:166.0
##  T      : 6107   Tuesday Lake  : 6107   Median :1997   Median :194.0
##  W      : 4188   West Long Lake: 4188   Mean   :1999   Mean   :194.3
##  E      : 3905   East Long Lake: 3905   3rd Qu.:2006   3rd Qu.:222.0
##  M      : 1234   Crampton Lake : 1234   Max.   :2016   Max.   :307.0
##  (Other): 1567   (Other)       : 1567
##    sampledate             depth       temperature_C   dissolvedOxygen
##  Min.   :1984-05-27   Min.   : 0.00   Min.   : 0.30   Min.   :  0.00
##  1st Qu.:1991-08-08   1st Qu.: 1.50   1st Qu.: 5.30   1st Qu.:  0.30
##  Median :1997-07-28   Median : 4.00   Median : 9.30   Median :  5.60
##  Mean   :1999-02-05   Mean   : 4.39   Mean   :11.81   Mean   :  4.97
```

```
##  3rd Qu.:2006-06-06   3rd Qu.: 6.50   3rd Qu.:18.70   3rd Qu.:  8.40
##  Max.   :2016-08-17   Max.   :20.00   Max.   :34.10   Max.   :802.00
##                                       NA's   :3858    NA's   :4039
##  irradianceWater     irradianceDeck                                comments
##  Min.   :   -0.337   Min.   :   1.5   DO Probe bad - Doesn't go to zero:  206
##  1st Qu.:   14.000   1st Qu.: 353.0   DO taken with Jones Lab Meter    :  162
##  Median :   65.000   Median : 747.0   NA's                             :38246
##  Mean   :  210.242   Mean   : 720.5
##  3rd Qu.:  265.000   3rd Qu.:1042.0
##  Max.   :24108.000   Max.   :8532.0
##  NA's   :14287       NA's   :15419
```

```
# 2
my_theme <- theme_bw(base_size = 12) + theme(axis.text = element_text(color = "black"),
    legend.position = "top", legend.justification = "center") +
    theme(plot.title = element_text(hjust = 0.5))
theme_set(my_theme)
```

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 tempearature recorded during July will change with depth across all lakes.

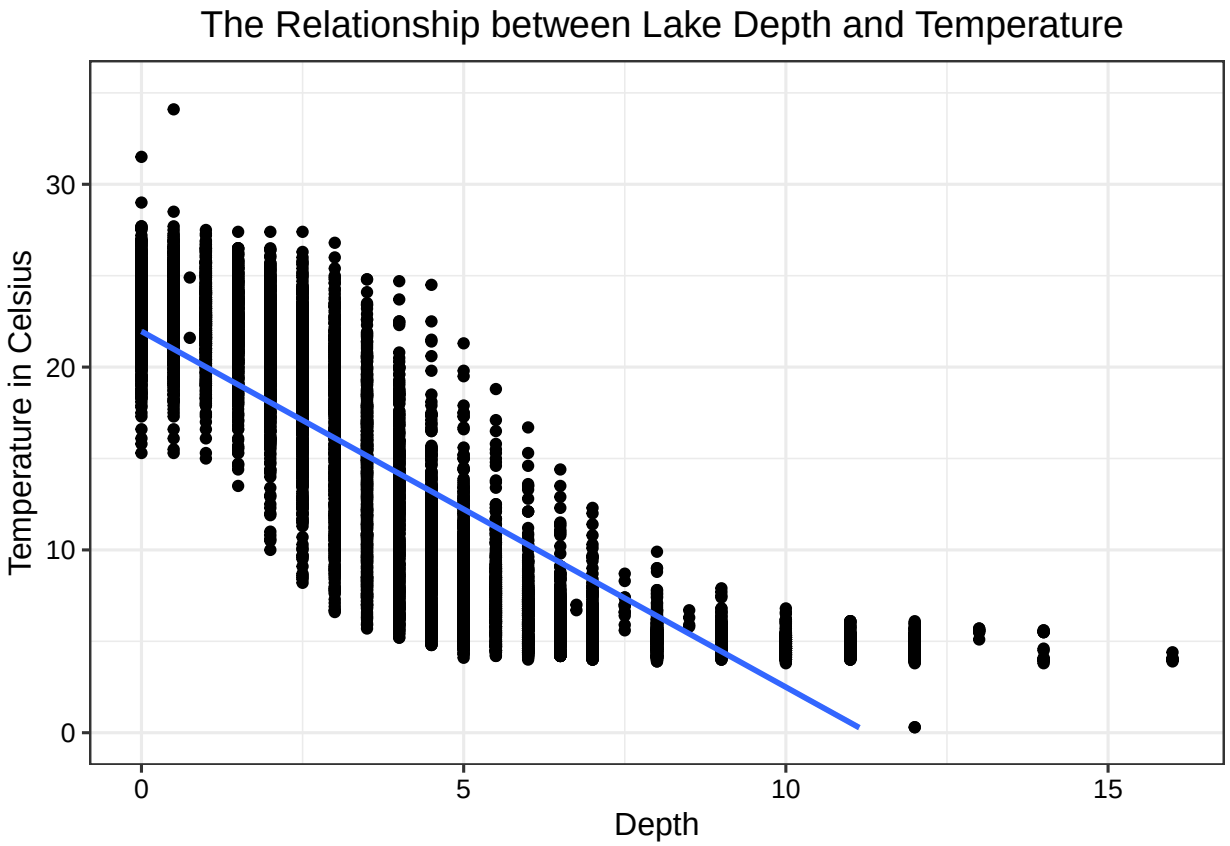
4. Wrangle your NTL-LTER dataset with a pipe function so that the records meet the following criteria:

- 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.

```
# 4

processed.lake.chemistry <- lake.chemistry %>%
    mutate(Month = month(sampledate)) %>%
    filter(Month == 7) %>%
    select(c(lakename, year4, daynum, depth, temperature_C)) %>%
    drop_na()  ## Subset for the month of July

# 5.

ggplot(processed.lake.chemistry, aes(x = depth, y = temperature_C)) +
    geom_point() + geom_smooth(method = lm) + ylim(0, 35) + ylab("Temperature in Celsius") +
    xlab("Depth") + ggtitle(" The Relationship between Lake Depth and Temperature")
```

```
## `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 graph suggests that temperature is inversely related to depth. As depth increases, temperature decreases. The distribution of points suggest that the relationship is not linear but more closely follows a curve.

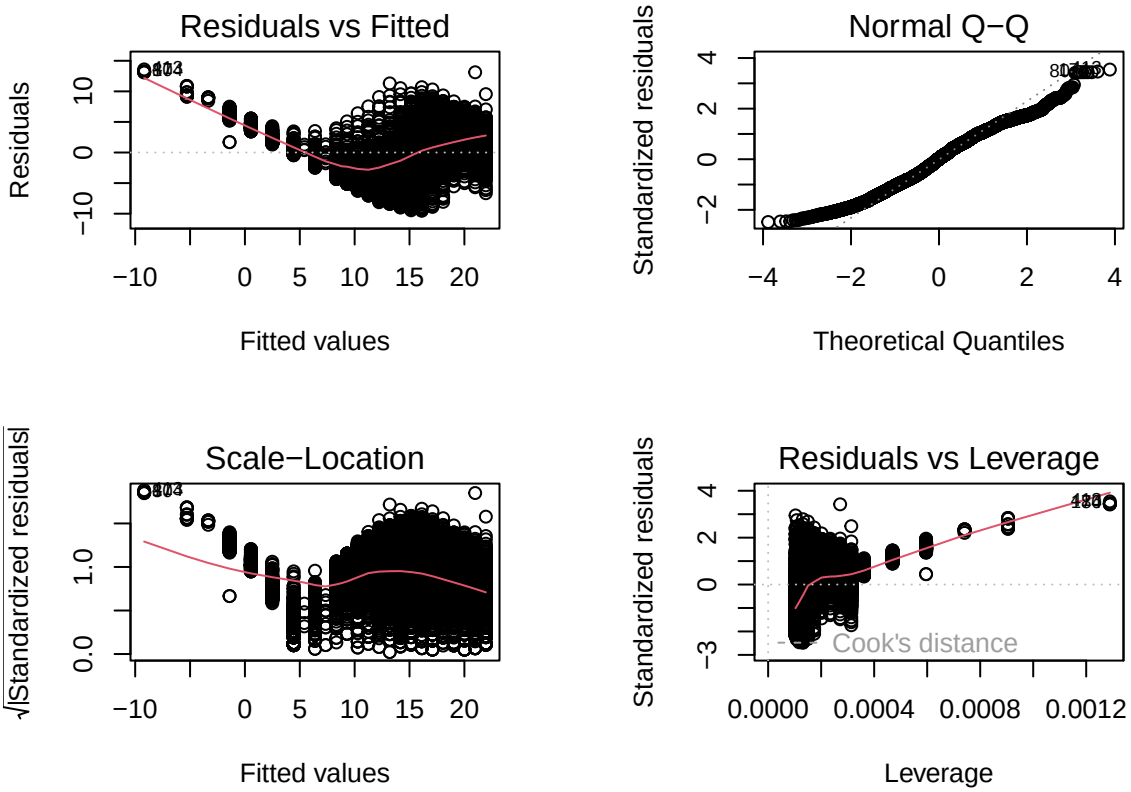
7. Perform a linear regression to test the relationship and display the results

```
# 7

lake.chemistry.mod <- lm(data = processed.lake.chemistry, temperature_C ~
    depth)

summary(lake.chemistry.mod)  #Display results
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = processed.lake.chemistry)
##
## 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(lake.chemistry.mod)  # Check the fit of the model
```



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: Based on the results on the model, 73.87% of the variability in temperature is explained by changes in depth. The degrees of freedom on which this finding is based is 9726. The p-value for this model is less than 0.001.

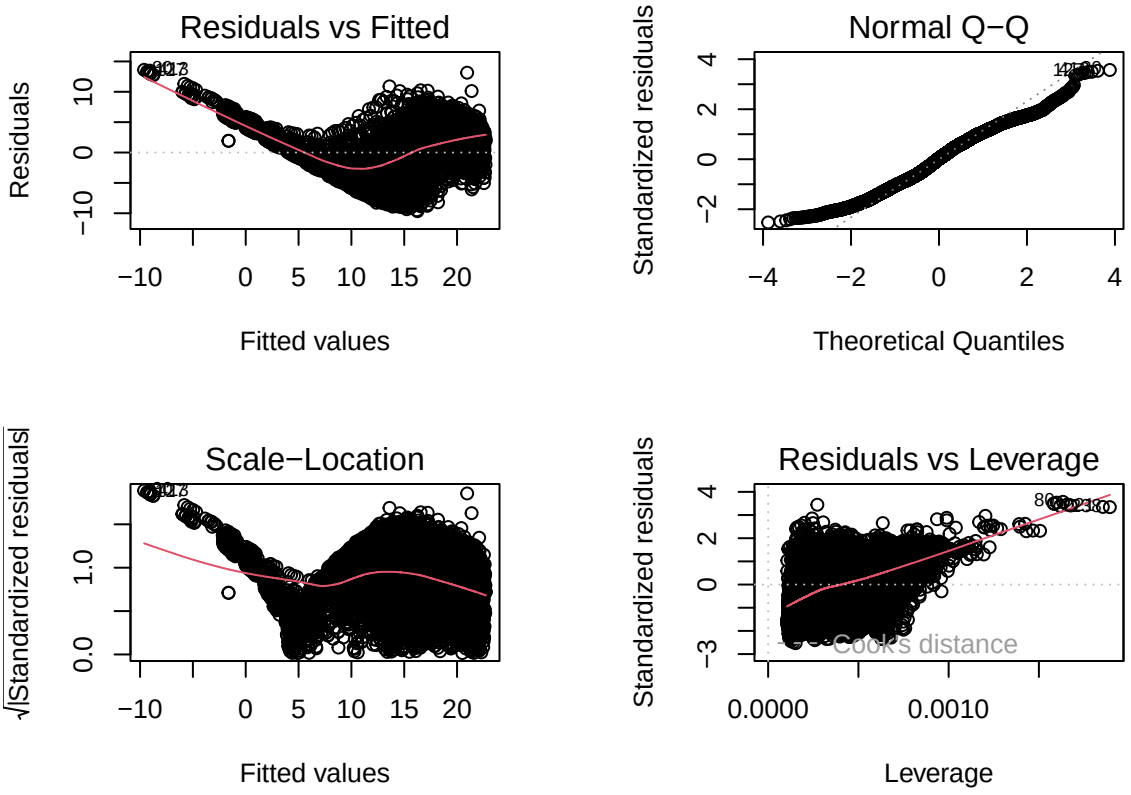
Interpretation: For every 1m change in depth, there is temperature decreases by -1.94621 Celsius.

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.

# Determine the set of explanatory variables best suited to
# predict temperature


lake.chemistry.mod.all.v <- lm(data = processed.lake.chemistry,
    temperature_C ~ year4 + daynum + depth)
# Model which includes all variables

par(mfrow = c(2, 2), mar = c(4, 4, 4, 4))
plot(lake.chemistry.mod.all.v)
```



```
summary(lake.chemistry.mod.all.v)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = processed.lake.chemistry)
##
## 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
```

```
step(lake.chemistry.mod.all.v)  ##Step Model
```

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

```
## AIC is the best when you remove none.

# 10.

lm(data = processed.lake.chemistry, temperature_C ~ year4 + daynum +
    depth)
```

```
##
## Call:
```

```
## lm(formula = temperature_C ~ year4 + daynum + depth, data = processed.lake.chemistry)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##    -8.57556      0.01134      0.03978     -1.94644
```

```
summary(lake.chemistry.mod.all.v)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = processed.lake.chemistry)
##
## 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 final set of explanatory variables that the AIC method suggests to predict temperature is year4, dayum and depth. This model explains 74.11% of the observed variance. This model is a slight improvemnet over the model using only depth as an explanatory 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.

## Formatted as ANOVA

lake.chemistry.anova <- aov(data = processed.lake.chemistry,
    temperature_C ~ lakename)
summary(lake.chemistry.anova)  ## ANOVA Model
```

```
##               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
```

```
## Formatted as Linear Model

lake.chemistry.anova2 <- lm(data = processed.lake.chemistry,
    temperature_C ~ lakename)
summary(lake.chemistry.anova2)  ## Linear Model
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = processed.lake.chemistry)
##
## 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: Given the null hypothesis is true, there is a significant difference in mean temperature among the lakes where the p-value is less than 0.001. Therefore, we can reject the null hypothesis in favor of the alternative that is a difference in mean lake temperatures.

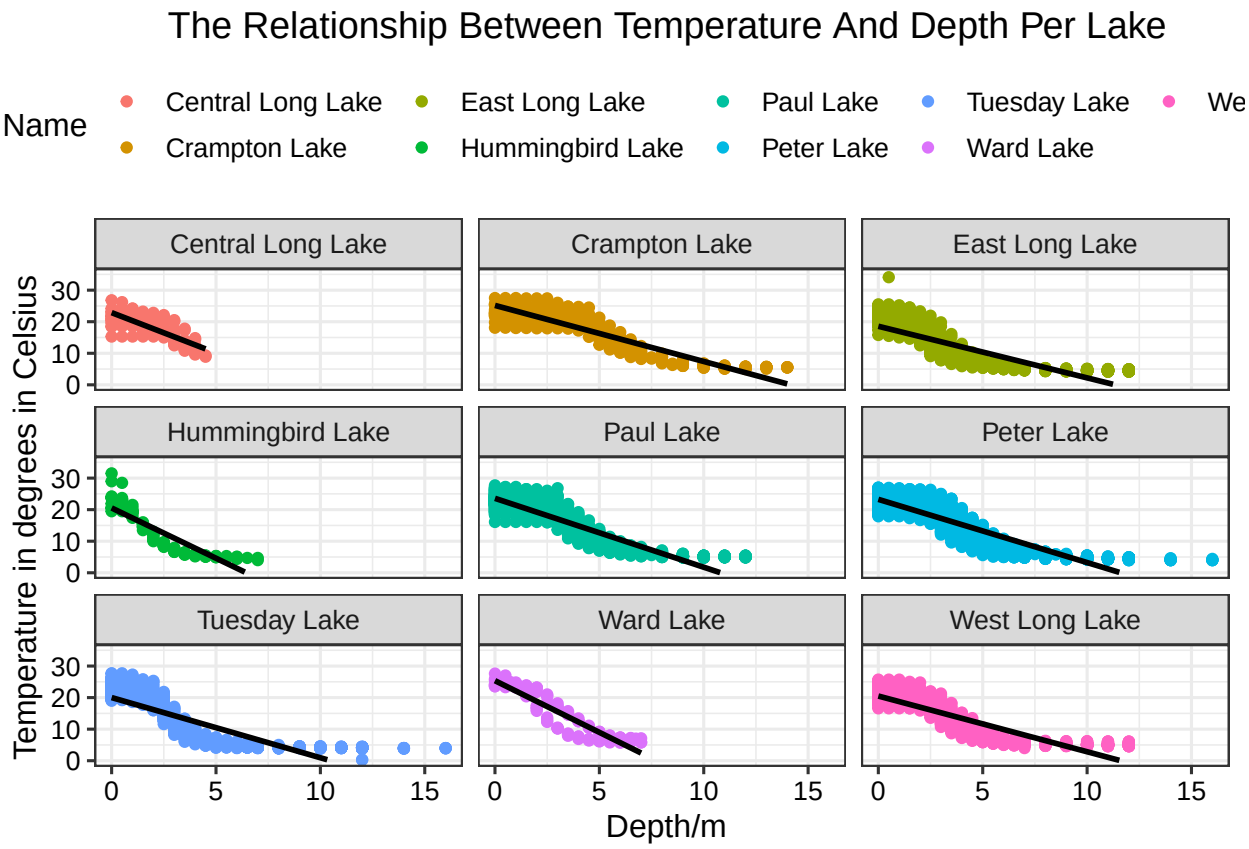
14. Create a graph that depicts temperature by depth, with a separate color for each lake. Add a geom_smooth (method = "lm", se = FALSE) for each lake. Make your points 50 % transparent. Adjust your y axis limits to go from 0 to 35 degrees. Clean up your graph to make it pretty.

```
# 14.

ggplot(processed.lake.chemistry, aes(x = depth, y = temperature_C,
    color = lakename)) + geom_point() + geom_smooth(method = "lm",
```

```
    color = "black", se = FALSE) + ylab("Temperature in degrees in Celsius") +
    xlab(" Depth/m") + ggtitle(" The Relationship Between Temperature And Depth Per Lake") +
    ylim(0, 35) + facet_wrap(vars(lakename), nrow = 3) + labs(color = "Lake Name")
```

```
## `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.chemistry.anova)
```

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

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: Based on the findings, the following lakes have the same mean temperature as Peter Lake, statisttically speaking: Ward Lake and Paul Lake. Based on these results, no lake has a mean temperature that is different from ALL other lakes another based on the adjusted p-values.

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 just looking at Peter and Paul Lake, we could use a two-sampled t-test.

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.lake.chemistry.July <- subset(processed.lake.chemistry,
    lakename %in% c("Crampton Lake", "Ward Lake"))

summary(Crampton.Ward.lake.chemistry.July)
```

```
##               lakename       year4          daynum          depth
##  Crampton Lake    :318   Min.   :1999   Min.   :183.0   Min.   : 0.000
##  Ward Lake        :116   1st Qu.:2004   1st Qu.:188.0   1st Qu.: 2.000
##  Central Long Lake:  0   Median :2005   Median :197.0   Median : 4.500
##  East Long Lake   :  0   Mean   :2006   Mean   :196.7   Mean   : 4.937
##  Hummingbird Lake :  0   3rd Qu.:2010   3rd Qu.:204.0   3rd Qu.: 7.000
##  Paul Lake        :  0   Max.   :2012   Max.   :211.0   Max.   :14.000
##  (Other)          :  0
##  temperature_C
##  Min.   : 5.00
##  1st Qu.: 7.40
##  Median :15.30
##  Mean   :15.11
##  3rd Qu.:22.38
##  Max.   :27.60
##
```

```
## Check that data was subsetted correctly


##############################################

# Testing the mean difference in temperature

t.test(data = Crampton.Ward.lake.chemistry.July, temperature_C ~
    lakename)
```

```
##
##  Welch Two Sample t-test
##
## data:  temperature_C by 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: Based on the results of the t-test, there is no significant difference between the mean temperature for Crampton Lake and Ward Lake, p- value is greater than 0.05. The means are the same statistically. We do not have sufficient evidence to reject the null hypothesis. This is synonymous with my response to question 16.