# Assignment 6: Generalized Linear Models

*Njeri Kara*

## OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics (ENV872L) on generalized linear models.

## Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Use the lesson as a guide. It contains code that can be modified to complete the assignment.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure to **answer the questions** in this assignment document. Space for your answers is provided in this document and is indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">". You should notice that the answer is highlighted in green by RStudio.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file. You will need to have the correct software installed to do this (see Software Installation Guide) Press the `Knit` button in the RStudio scripting panel. This will save the PDF output in your Assignments folder.
6. After Knitting, please submit the completed exercise (PDF file) to the dropbox in Sakai. Please add your last name into the file name (e.g., "Salk_A06_GLMs.pdf") prior to submission.

The completed exercise is due on Tuesday, 26 February, 2019 before class begins.

## Set up your session

1. Set up your session. Upload the EPA Ecotox dataset for Neonicotinoids and the NTL-LTER raw data file for chemistry/physics.

```
#1
#Setting the working directory
setwd("C:/Users/jerik/OneDrive/Documents/Spring 2019 Semenster/Environmental Data Analytics/EDA_R_Work/I
#Confirming that it is the correct working directory
getwd()
```

```
## [1] "C:/Users/jerik/OneDrive/Documents/Spring 2019 Semenster/Environmental Data Analytics/EDA_R_Work,
```

```
#Loading necessary packages
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

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

```
library(lubridate)
```

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

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(knitr)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
#Uploading the required dataset.
ECOTOX.Neonicotinoids.Mortality.raw.D <-
  read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Mortality_raw.csv")

NTL.Lake.Chem.Phy.Raw <-
  read.csv("./Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv")

#Exploring the datasets
str(ECOTOX.Neonicotinoids.Mortality.raw.D)
```

```
## 'data.frame':    1283 obs. of  13 variables:
##  $ CAS.No.          : int  138261413 111988499 138261413 138261413 111988499 111988499 111988499 111
##  $ Chemical.Name    : Factor w/ 9 levels "Acetamiprid",..: 4 8 4 4 8 8 8 8 4 4 ...
##  $ Species.Name     : Factor w/ 172 levels "Acipenser transmontanus",..: 54 86 54 43 54 54 54 54 43 
##  $ Common.Name      : Factor w/ 124 levels "Alderfly","Alfalfa Plant Bug",..: 68 97 68 68 68 68 68 68
##  $ Effect           : Factor w/ 1 level "Mortality": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Measurement      : Factor w/ 1 level "Mortality": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Endpoint         : Factor w/ 23 levels "EC10","EC50",..: 5 23 9 5 5 5 5 9 9 20 ...
##  $ Dur..Std.        : num  28 7 28 28 21 28 14 28 28 4 ...
##  $ Conc..Type       : Factor w/ 3 levels "Active ingredient",..: 2 1 2 2 1 1 1 1 2 1 ...
##  $ Conc..Mean..Std. : num  0.000041 0.00007 0.000195 0.000235 0.00024 0.00027 0.0003 0.0003 0.000316
##  $ Conc..Units..Std.: Factor w/ 16 levels "AI mg/kg bdwt",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Pub..Year        : int  2013 2017 2013 2013 2016 2016 2016 2016 2013 1992 ...
##  $ Citation         : Factor w/ 198 levels "Aaen,S.M., L.A. Hamre, and T.E. Horsberg. A Screening of
```

```
str(NTL.Lake.Chem.Phy.Raw )
```

```
## 'data.frame':    38614 obs. of  11 variables:
##  $ lakeid         : Factor w/ 9 levels "C","E","H","L",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ lakename       : Factor w/ 9 levels "Central Long Lake",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ year4          : int  1984 1984 1984 1984 1984 1984 1984 1984 1984 1984 ...
##  $ daynum         : int  148 148 148 148 148 148 148 148 148 148 ...
##  $ sampledate     : Factor w/ 1712 levels "10/1/07","10/1/93",..: 134 134 134 134 134 134 134 134 13
##  $ depth          : num  0 0.25 0.5 0.75 1 1.5 2 3 4 5 ...
##  $ temperature_C  : num  14.5 NA NA NA 14.5 NA 14.2 11 7 6.1 ...
##  $ dissolvedOxygen: num  9.5 NA NA NA 8.8 NA 8.6 11.5 11.9 2.5 ...
##  $ irradianceWater: num  1750 1550 1150 975 870 610 420 220 100 34 ...
##  $ irradianceDeck : num  1620 1620 1620 1620 1620 1620 1620 1620 1620 1620 ...
##  $ comments       : Factor w/ 2 levels "DO Probe bad - Doesn't go to zero",..: NA NA NA NA NA NA NA
```

```
#changing date variable to date format for
NTL.Lake.Chem.Phy.Raw$sampledate <- as.Date(NTL.Lake.Chem.Phy.Raw $sampledate,format = "%m/%d/%y")
```

2. Build a ggplot theme and set it as your default theme.

```
#2
#building a theme
NK.theme <- theme_light(base_size = 12) +
  theme(plot.background = element_rect(fill = "grey97"),
        panel.grid.major =element_line(linetype = "dotted"),
        panel.grid.minor = element_line(linetype = "dotted"), text=element_text(size = 14, color = "bla

#setting it as my default theme
theme_set(NK.theme)
```

## Neonicotinoids test

Research question: Were studies on various neonicotinoid chemicals conducted in different years?

3. Generate a line of code to determine how many different chemicals are listed in the Chemical.Name column.

```
#3
levels(ECOTOX.Neonicotinoids.Mortality.raw.D$Chemical.Name)
```

```
## [1] "Acetamiprid"  "Clothianidin" "Dinotefuran"  "Imidacloprid"
## [5] "Imidaclothiz" "Nitenpyram"   "Nithiazine"   "Thiacloprid"
## [9] "Thiamethoxam"
```

4. Are the publication years associated with each chemical well-approximated by a normal distribution? Run the appropriate test and also generate a frequency polygon to illustrate the distribution of counts for each year, divided by chemical name. Bonus points if you can generate the results of your test from a pipe function. No need to make this graph pretty.

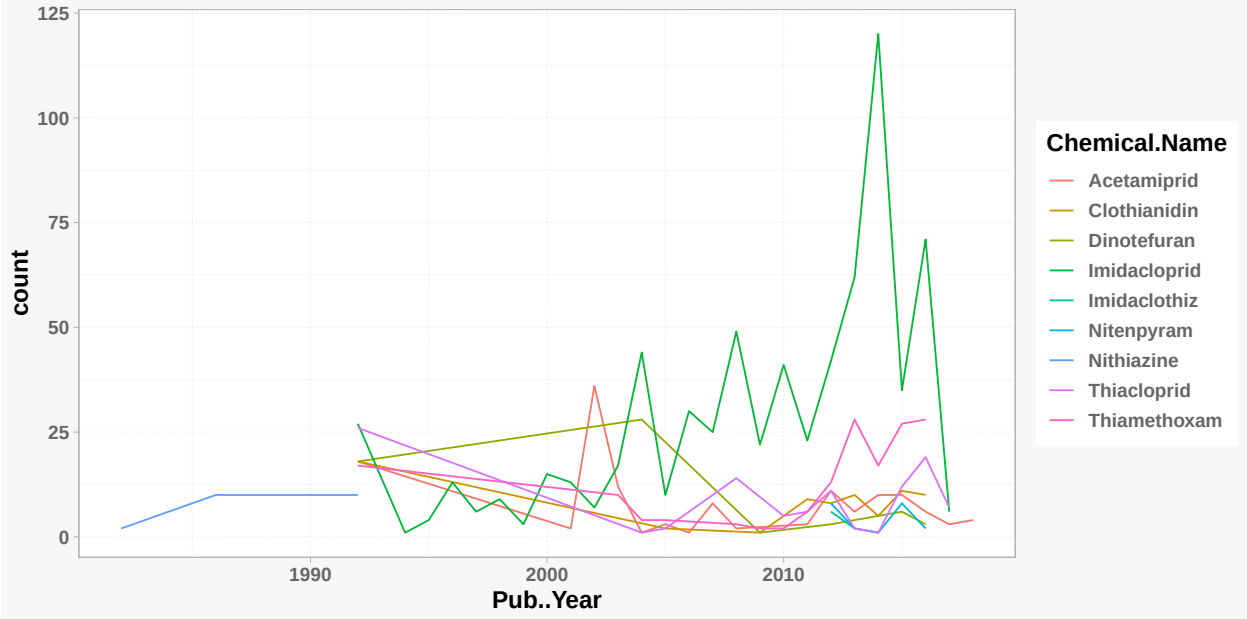
```
#4
chem.names <- unique(ECOTOX.Neonicotinoids.Mortality.raw.D$Chemical.Name)

for (i in chem.names){
  Norm.test.all <- ECOTOX.Neonicotinoids.Mortality.raw.D %>%
  filter(Chemical.Name == i) %>%
  pull(Pub..Year) %>%
  shapiro.test()
  print(paste0("Shapiro test for years associated with chemical ", i))
  print(Norm.test.all)
}
```

```
## [1] "Shapiro test for years associated with chemical Imidacloprid"
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.88178, p-value < 2.2e-16
##
## [1] "Shapiro test for years associated with chemical Thiacloprid"
##
##  Shapiro-Wilk normality test
```

```
##
## data:  .
## W = 0.7669, p-value = 1.118e-11
##
## [1] "Shapiro test for years associated with chemical Thiamethoxam"
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.7071, p-value < 2.2e-16
##
## [1] "Shapiro test for years associated with chemical Acetamiprid"
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.90191, p-value = 5.706e-08
##
## [1] "Shapiro test for years associated with chemical Clothianidin"
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.69577, p-value = 4.287e-11
##
## [1] "Shapiro test for years associated with chemical Dinotefuran"
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.82848, p-value = 8.83e-07
##
## [1] "Shapiro test for years associated with chemical Nitenpyram"
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.79592, p-value = 0.0005686
##
## [1] "Shapiro test for years associated with chemical Nithiazine"
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.75938, p-value = 0.0001235
##
## [1] "Shapiro test for years associated with chemical Imidaclothiz"
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.68429, p-value = 0.00093
```

```
#Frequency polygon graph
ggplot(ECOTOX.Neonicotinoids.Mortality.raw.D, aes(x = Pub..Year, color = Chemical.Name)) +
  geom_freqpoly(stat = "count") +
  NK.theme
```




The publication years associated with each chemical are not well approxiamted by a normal distribution. The shapiro tests of publication year for each chemical all have a p-value less than 0.5 meaning we reject the Ho hypothis that the publication years have a normal distribution.

5. Is there equal variance among the publication years for each chemical? Hint: var.test is not the correct function.

```
#5
#testing for equal variance of pairings
bartlett.test(ECOTOX.Neonicotinoids.Mortality.raw.D$Pub..Year ~ ECOTOX.Neonicotinoids.Mortality.raw.D$Cl
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  ECOTOX.Neonicotinoids.Mortality.raw.D$Pub..Year by ECOTOX.Neonicotinoids.Mortality.raw.D$Chem
## Bartlett's K-squared = 139.59, df = 8, p-value < 2.2e-16
```

No there is no equal variance among the publication years for each chemical. When we run the bartlett test comparing variance of publication year across the different chemicals, the p_value of the test is < 2.2e-16 indicating we reject the Ho hypothesis of the test that the variances are equal.

6. Based on your results, which test would you choose to run to answer your research question?

   Based on results to answer the research quastion - were studies on various neonicotinoid chemicals conducted in different years - the test that should be used is the Kruskal-Wallis test. This is because this test is the Non-parametric equivalent of the ANOVA test.

7. Run this test below.

```
#7
#Kruskal test
```

```
ECOTOX.yr.chem.test <- kruskal.test(ECOTOX.Neonicotinoids.Mortality.raw.D$Pub..Year ~ ECOTOX.Neonicotino
ECOTOX.yr.chem.test
```

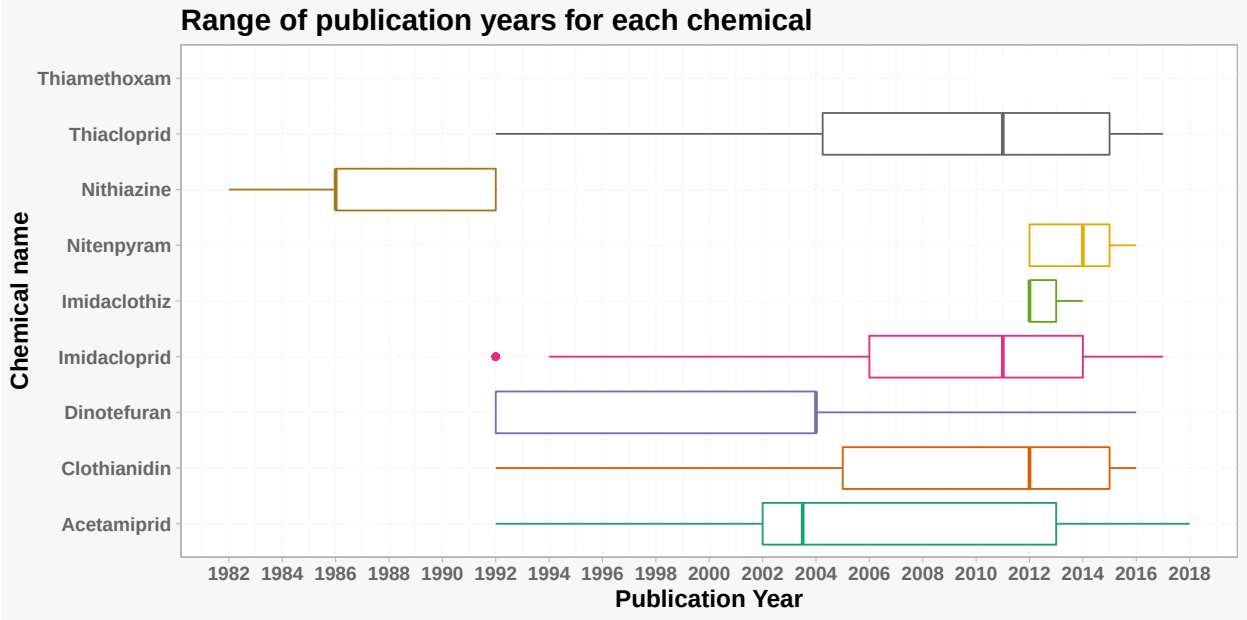
```
##
##  Kruskal-Wallis rank sum test
##
## data:  ECOTOX.Neonicotinoids.Mortality.raw.D$Pub..Year by ECOTOX.Neonicotinoids.Mortality.raw.D$Chemi
## Kruskal-Wallis chi-squared = 134.15, df = 8, p-value < 2.2e-16
```

8. Generate a boxplot representing the range of publication years for each chemical. Adjust your graph to make it pretty.

```
#8
Ecotox.plot.Q8 <- ggplot(ECOTOX.Neonicotinoids.Mortality.raw.D,
                         aes(x = Chemical.Name, y = Pub..Year)) +
  geom_boxplot(aes(color = Chemical.Name)) +
  coord_flip() +
  theme(legend.position = "none") +
  scale_color_brewer(palette="Dark2") +
  ggtitle("Range of publication years for each chemical") +
  scale_y_continuous(breaks = seq(min(ECOTOX.Neonicotinoids.Mortality.raw.D$Pub..Year), max(ECOTOX.Neoni
  xlab("Chemical name") +
  ylab("Publication Year")

print(Ecotox.plot.Q8)
```

```
## Warning in RColorBrewer::brewer.pal(n, pal): n too large, allowed maximum for palette Dark2 is 8
## Returning the palette you asked for with that many colors
```




9. How would you summarize the conclusion of your analysis? Include a sentence summarizing your findings and include the results of your test in parentheses at the end of the sentence.

   ANSWER: There is a significant difference studies on various neonicotinoid chemicals conducted in different year (Kruskal-Wallis test; Kruskal-Wallis chi-squared = 134.15, df = 8, p-value < 2.2e-16)

## NTL-LTER test

Research question: What is the best set of predictors for lake temperatures in July across the monitoring period at the North Temperate Lakes LTER?

11. Wrangle your NTL-LTER dataset with a pipe function so that it contains only the following criteria:

- Only dates in July (hint: use the daynum column). No need to consider leap years.
- Only the columns: lakename, year4, daynum, depth, temperature_C
- Only complete cases (i.e., remove NAs)

```
#11
NTL.df <- NTL.Lake.Chem.Phy.Raw  %>%
  filter(daynum %in% c(182:212)) %>%
  select(lakename,year4,daynum,depth,temperature_C) %>%
  na.omit()
```

12. Run an AIC to determine what set of explanatory variables (year4, daynum, depth) is best suited to predict temperature. Run a multiple regression on the recommended set of variables.

```
#12
Q12.AIC.test <- lm(data = NTL.df, temperature_C ~ year4 +
                     daynum + depth)
step(Q12.AIC.test)
```

```
## Start:  AIC=26016.31
## temperature_C ~ year4 + daynum + depth
##
##          Df Sum of Sq    RSS   AIC
## <none>                 141118 26016
## - year4   1        80 141198 26020
## - daynum  1      1333 142450 26106
## - depth   1    403925 545042 39151
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL.df)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##    -6.45556      0.01013      0.04134     -1.94726
```

```
Q12.model <- lm(data = NTL.df, temperature_C ~ year4 +
                  daynum + depth)
summary(Q12.model)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6517 -2.9937  0.0855  2.9692 13.6171
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -6.455560   8.638808   -0.747   0.4549
## year4        0.010131   0.004303    2.354   0.0186 *
```

```
## daynum       0.041336   0.004315    9.580   <2e-16 ***
## depth       -1.947264   0.011676 -166.782   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.811 on 9718 degrees of freedom
## Multiple R-squared:  0.7417, Adjusted R-squared:  0.7417
## F-statistic:  9303 on 3 and 9718 DF,  p-value: < 2.2e-16
```

13. What is the final linear equation to predict temperature from your multiple regression? How much of the observed variance does this model explain?

    ANSWER: The final linear equation to predict temperature is:

$$temperature\_c = -6.455560 + (0.010131 * year4) + (0.041336 * daynum) + (1.947264 * depth)$$

,

$$(R^2 = 0.7417, F - statistic : 9303 on 3 and 9718 DF, p < 2.2e - 16)$$

> (lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL.df). The model explains 74.17% of the variation in temperature observations.

14. Run an interaction effects ANCOVA to predict temperature based on depth and lakename from the same wrangled dataset.

```r
#14
# interaction effects ANCOVA
Q14.ancova.int <- lm(data = NTL.df, temperature_C ~ depth * lakename )
summary(Q14.ancova.int)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth * lakename, data = NTL.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6455 -2.9133 -0.2879  2.7567 16.3606
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      22.9455     0.5861  39.147  < 2e-16 ***
## depth                            -2.5820     0.2411 -10.711  < 2e-16 ***
## lakenameCrampton Lake             2.2173     0.6804   3.259  0.00112 **
## lakenameEast Long Lake           -4.3884     0.6191  -7.089 1.45e-12 ***
## lakenameHummingbird Lake         -2.4126     0.8379  -2.879  0.00399 **
## lakenamePaul Lake                 0.6105     0.5983   1.020  0.30754
## lakenamePeter Lake                0.2998     0.5970   0.502  0.61552
## lakenameTuesday Lake             -2.8932     0.6060  -4.774 1.83e-06 ***
## lakenameWard Lake                 2.4180     0.8434   2.867  0.00415 **
## lakenameWest Long Lake           -2.4663     0.6168  -3.999 6.42e-05 ***
## depth:lakenameCrampton Lake       0.8058     0.2465   3.268  0.00109 **
## depth:lakenameEast Long Lake      0.9465     0.2433   3.891  0.00010 ***
## depth:lakenameHummingbird Lake   -0.6026     0.2919  -2.064  0.03903 *
## depth:lakenamePaul Lake           0.4022     0.2421   1.662  0.09664 .
## depth:lakenamePeter Lake          0.5799     0.2418   2.398  0.01649 *
## depth:lakenameTuesday Lake        0.6605     0.2426   2.723  0.00648 **
## depth:lakenameWard Lake          -0.6930     0.2862  -2.421  0.01548 *
```

```
## depth:lakenameWest Long Lake     0.8154     0.2431   3.354  0.00080 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.471 on 9704 degrees of freedom
## Multiple R-squared:  0.7861, Adjusted R-squared:  0.7857
## F-statistic:  2097 on 17 and 9704 DF,  p-value: < 2.2e-16
```

```
summary.aov(Q14.ancova.int)
```

```
##                   Df Sum Sq Mean Sq  F value Pr(>F)
## depth              1 403868  403868 33525.96 <2e-16 ***
## lakename           8  20949    2619   217.37 <2e-16 ***
## depth:lakename     8   4687     586    48.64 <2e-16 ***
## Residuals       9704 116899      12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

15. Is there an interaction between depth and lakename? How much variance in the temperature observations does this explain?

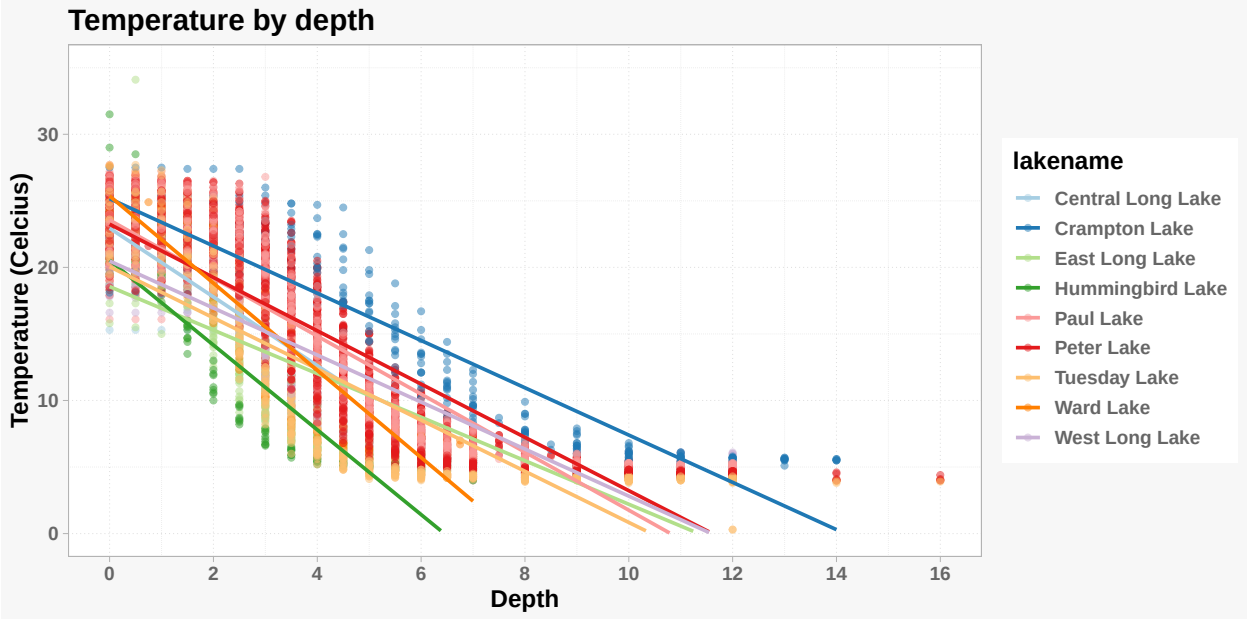
    ANSWER: Yes there is the interaction between depth and lakename has a significant effect in temperature (Adjusted R-squared: 0.7857 F-statistic: 2097 on 17 and 9704 DF, p-value: < 2.2e-16). This interaction explains 78.57% of the variance in temperature observations.

16. 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.

```
#16
Q16.plot <- ggplot(NTL.df, aes(x = depth, y = temperature_C, color = lakename)) +
  geom_point(alpha=0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  ylim(0,35) +
  scale_color_brewer(palette="Paired") +
  ggtitle("Temperature by depth") +
  scale_x_continuous(breaks = seq(min(NTL.df$depth), max(NTL.df$depth), by = 2)) + xlab("Depth") +
  ylab("Temperature (Celcius)")

print(Q16.plot)
```

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


Temperature by depth
Temperature (Celcius)
30
20
10
0
0
2
4
6
8
10
12
14
16
Depth
lakename
Central Long Lake
Crampton Lake
East Long Lake
Hummingbird Lake
Paul Lake
Peter Lake
Tuesday Lake
Ward Lake
West Long Lake