# Assignment 6: Generalized Linear Models

*Gaby Garcia*

## 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.

## 1.

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

```
setwd("~/Desktop/Environmental Data Analytics/Environmental_Data_Analytics/Data/Raw")
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.1.0     v purrr   0.3.0
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

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

```
library(tidyr)
library(ggplot2)
library(viridis)
```

```
## Loading required package: viridisLite
```

```
library(RColorBrewer)
library(colormap)
library(lubridate)
```

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

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

```
library(stats)

Neonicotinoids<-read.csv("ECOTOX_Neonicotinoids_Mortality_raw.csv")

PeterPaul.chem.nutrients <- read.csv("NTL-LTER_Lake_ChemistryPhysics_Raw.csv")
```

## 2.

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

```
library(ggthemes)
gabytheme <- theme_bw(base_size = 14) +
  theme(plot.title=element_text(face="bold", size="20", color="IndianRed4", hjust=0.5),
        axis.title=element_text(face="bold.italic", size=11, color="black"),
axis.text = element_text(face="bold", size=10, color = "black"),
panel.background=element_rect(fill="gray96", color="darkblue"),
panel.border = element_rect(color = "black", size = 2),
legend.position = "top", legend.background = element_rect(fill="white", color="black"),
            legend.key = element_rect(fill="transparent", color="NA"))
```

#### Set gabytheme as my default theme

```
theme_set(gabytheme)
```

# 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.

```
levels(Neonicotinoids$Chemical.Name)
```

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

### Shapiro.Test to Test Assumption of Normality

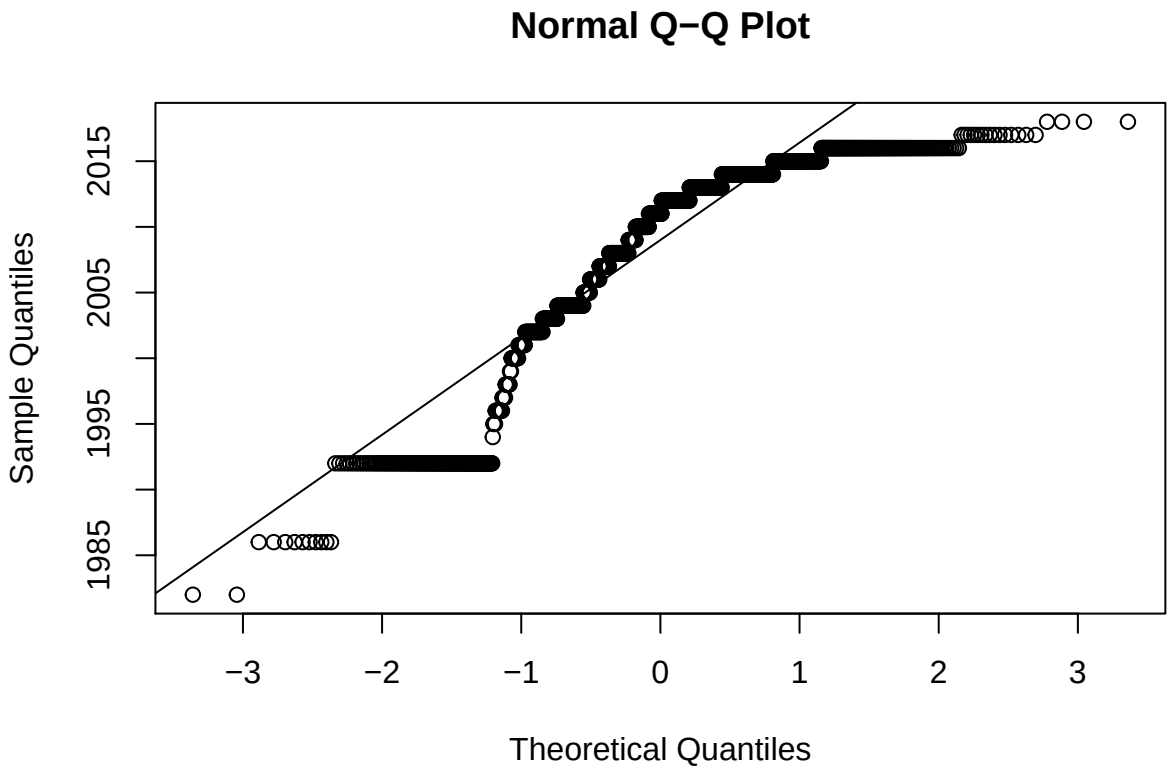
```
shapiro.test(Neonicotinoids$Pub..Year)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Neonicotinoids$Pub..Year
## W = 0.85472, p-value < 2.2e-16
```

We can reject the null hypothesis that the Publication years variable is distributed normally. (Test: Shapiro Wilks, W=0.854, p<2.2e-16)

**QQPlot shows that data is definitely not normally distributed!**

```
qqnorm(Neonicotinoids$Pub..Year)
qqline(Neonicotinoids$Pub..Year)
```



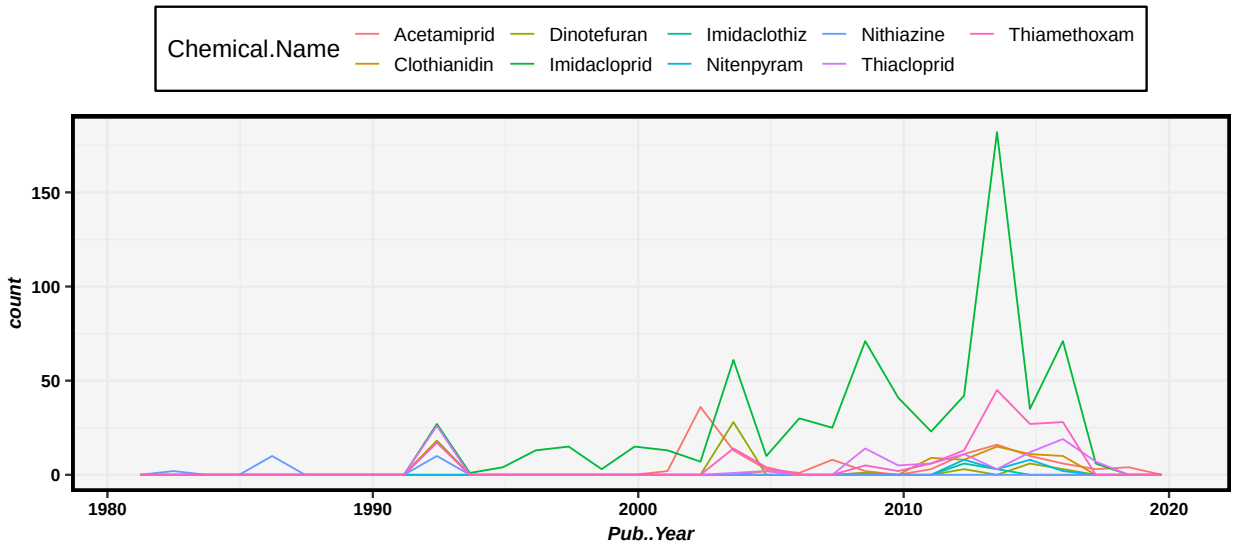

## 4.

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.

```
library(scales)

PubYearPlot <-
  ggplot(Neonicotinoids) +
  geom_freqpoly(aes(x =Pub..Year, color = Chemical.Name))

print(PubYearPlot)
```




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

**Use bartlett.test for non-normal distributions(var.test is for normallly distributed populations)**

```
bartlett.test(Neonicotinoids$Pub..Year~Neonicotinoids$Chemical.Name)
```

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

**Results: (Bartlett Test: K-Squared=139.59, df=8, p<2.2e-16). Based on the results of the Bartlett test, there is not equal variance among the publication years because the p-value of the test is <0.05, which means that we reject the null hypothesis that the variances are the same amongst the publication years for each chemical.**

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

ANSWER: The research question is "Were studies on various neonicotinoid chemicals conducted in different years?" We have a Continuous response variable of Publication Year, which is an integer, and we have one categorical explanatory variable with 9 levels, which is the Chemical Name variable. I would run a Kruskal-Wallis Test for this research question. The Kruskal Wallis

test is the non-parametric counterpart to the one-way ANOVA. The One-way ANOVA compares the means of the samples or groups in order to make inferences about the population means; there is only one independent variable or factor. Because our dependent variable has a non-normal distribution, we should use the Kruskal-Wallis instead of a regular one-way ANOVA.

## 7.

Run this test below.

```
NeoANOVA<- kruskal.test(Neonicotinoids$Pub..Year ~ Neonicotinoids$Chemical.Name)

NeoANOVA
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Neonicotinoids$Pub..Year by Neonicotinoids$Chemical.Name
## 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.

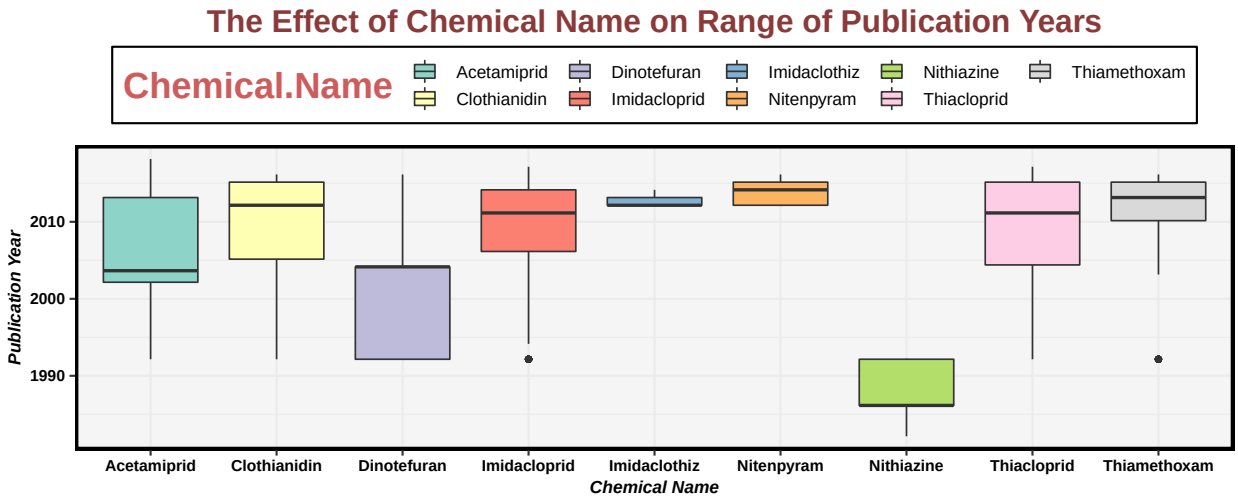
```
library(dplyr)
Neonicotinoids$Pub..Year<-as.factor(Neonicotinoids$Pub..Year)
Neonicotinoids$Pub..Year<-as.Date(Neonicotinoids$Pub..Year, format="%Y")
```

```
Neonicotinoids<- mutate(Neonicotinoids, year = year(Pub..Year))
Neonicotinoids$year<-as.factor(Neonicotinoids$year)
Neonicotinoids$year<-as.Date(Neonicotinoids$year, format="%Y")
Neonicotinoids<- separate(Neonicotinoids, year, c("Y", "m", "d"))
```

```
library(scales)

NeoPlot<-ggplot(Neonicotinoids, aes(x=Chemical.Name, y=Pub..Year, fill=Chemical.Name)) +
  geom_boxplot() +

labs(title="The Effect of Chemical Name on Range of Publication Years",
     x="Chemical Name", y="Publication Year") +
theme(legend.title = element_text(colour="IndianRed", size=23, face="bold"))+
 scale_fill_brewer(palette="Set3")   #use scale_fill_brewer for boxplots
print(NeoPlot)
```




# 9

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: Based on the results of my statistical tests and data visualization, the publication years for each of the 9 chemicals do not follow a normal distribution. There is a significant difference between the publication years across chemical types. (Kruskal-Wallis Non-Parametric Test, df=8, p<2.2e-16, Kruskal Wallis Chi Squared=134.15)

## 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

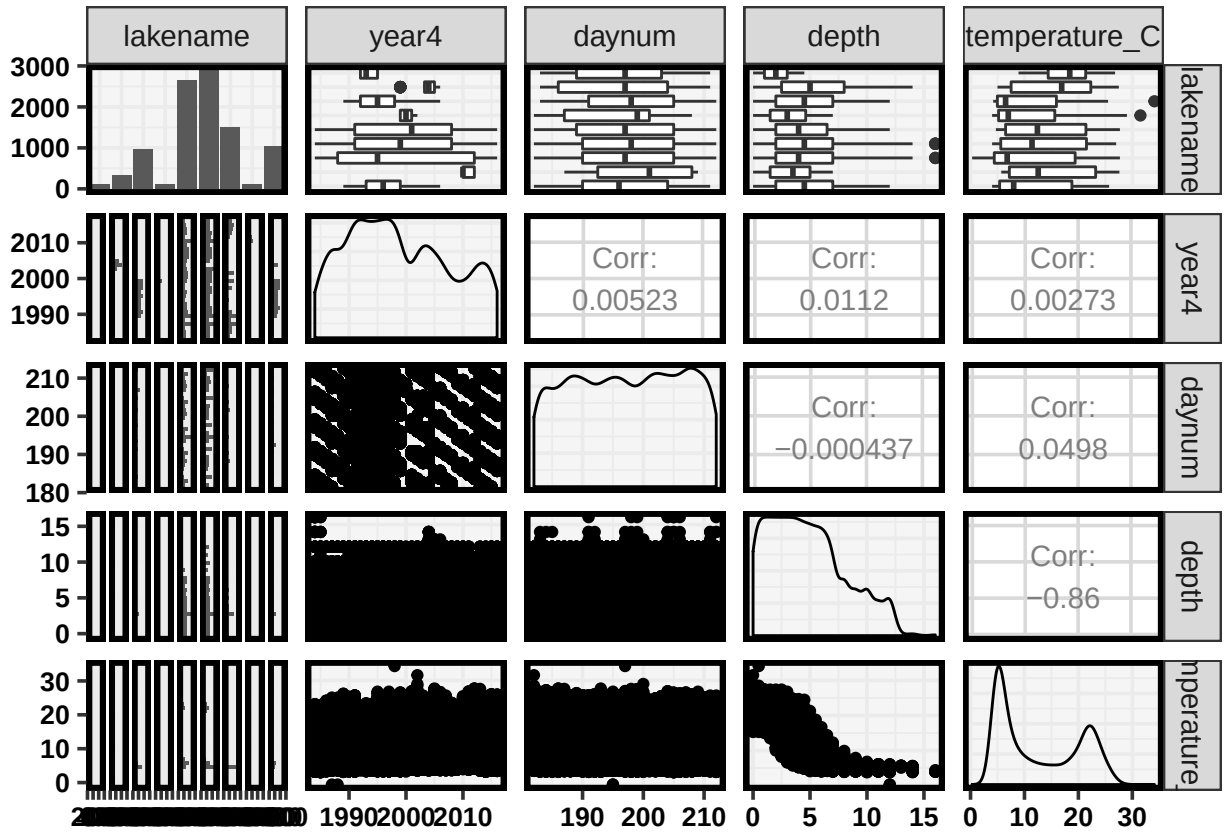
11. Wrangle your NTL-LTER dataset with a pipe function so that it contains only the following criteria:

- 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)

```
PeterPaulJulyFiltered<-filter(PeterPaul.chem.nutrients, daynum %in% c(182:212))%>%select(lakename, year4
```

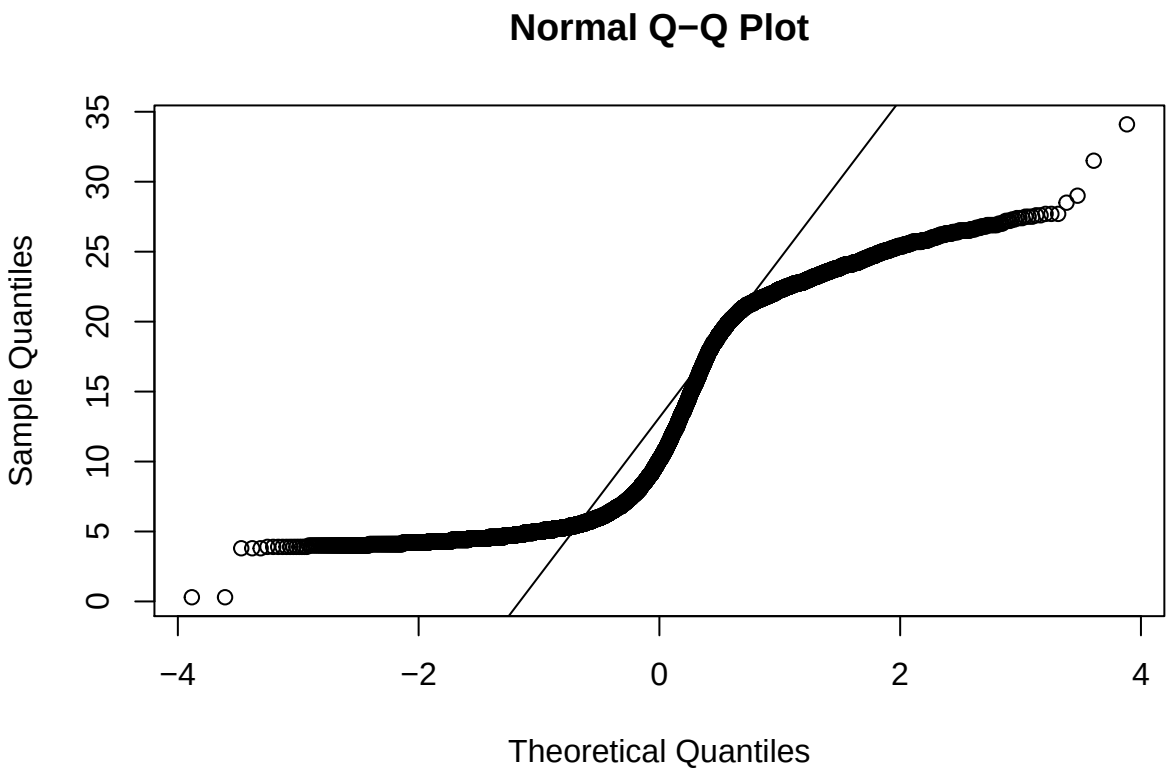
## Visualize Data Relationships

```
library(GGally)
ggpairs(PeterPaulJulyFiltered)
```



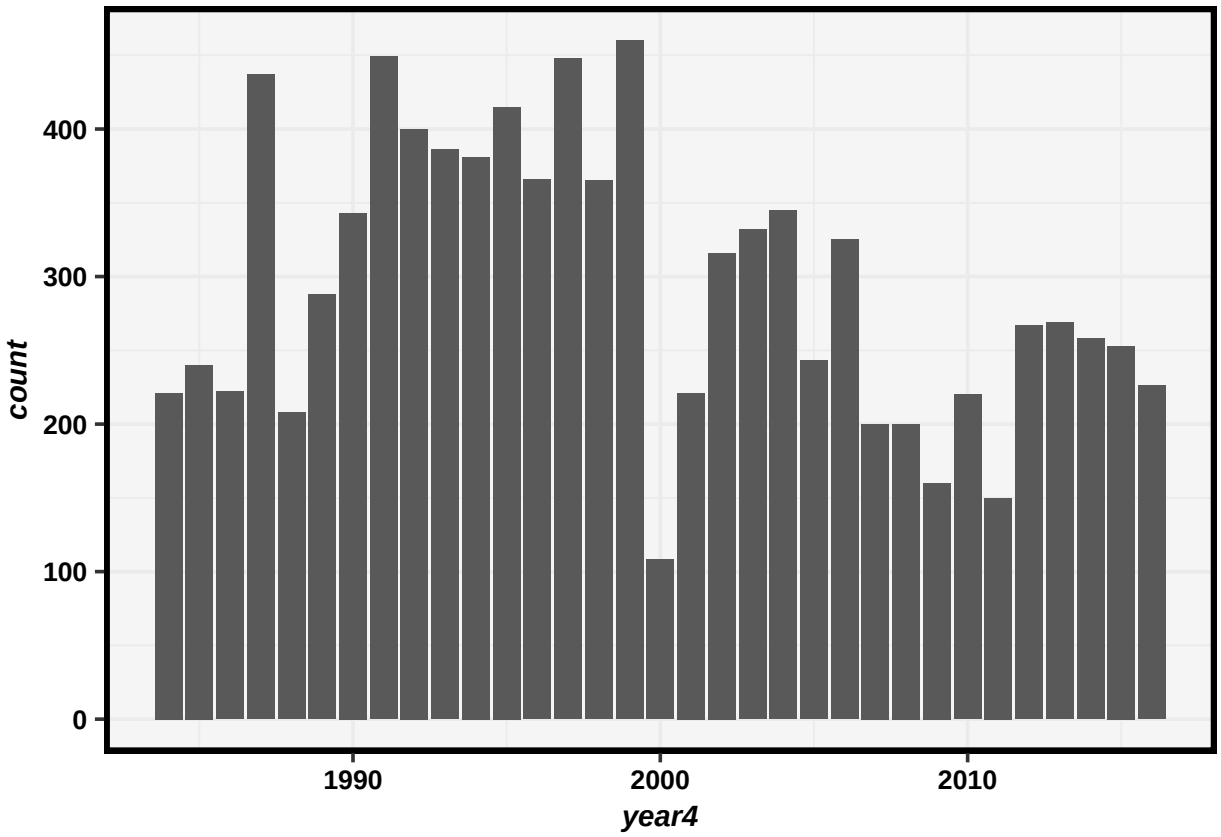

### Use QQNorm and QQline on Dependant Variable of Temperature

```
qqnorm(PeterPaulJulyFiltered$temperature_C)
qqline(PeterPaulJulyFiltered$temperature_C)
```

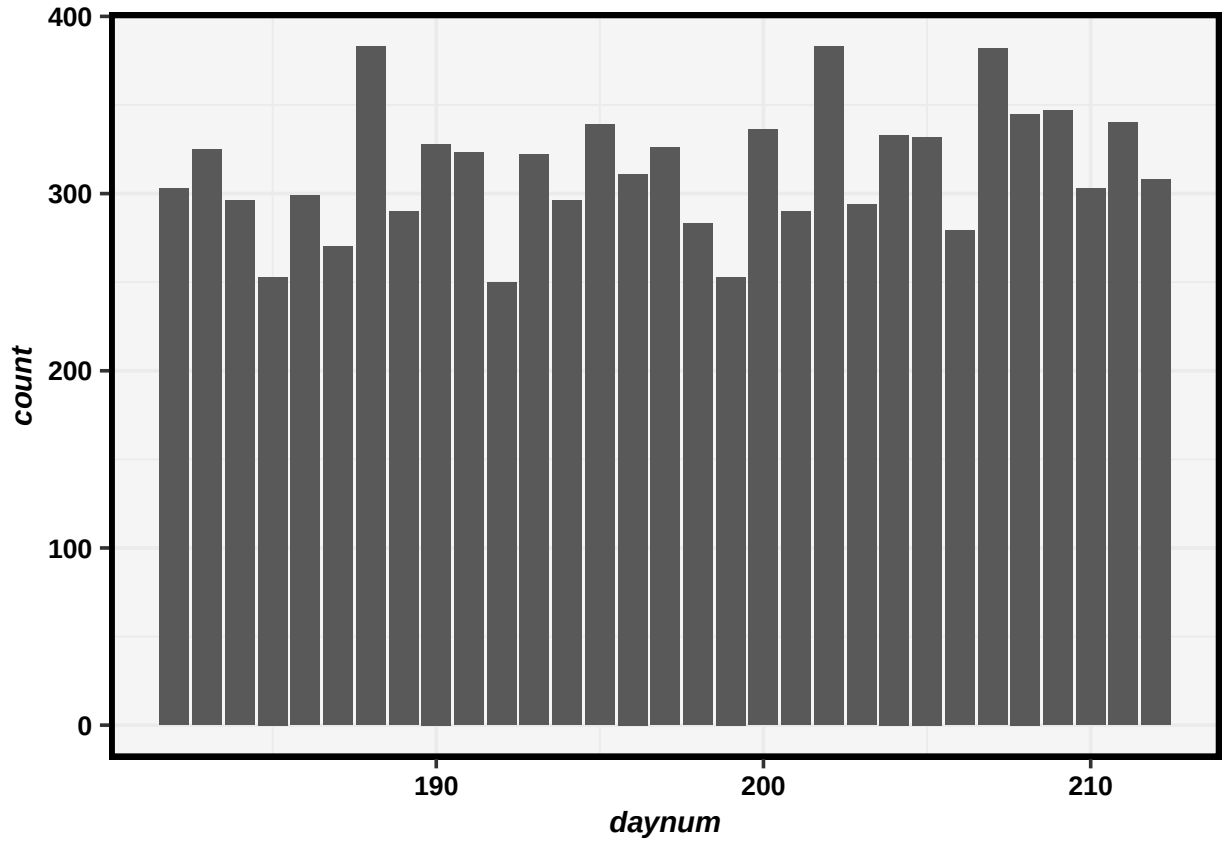



###Dependent variable is not normally distributed

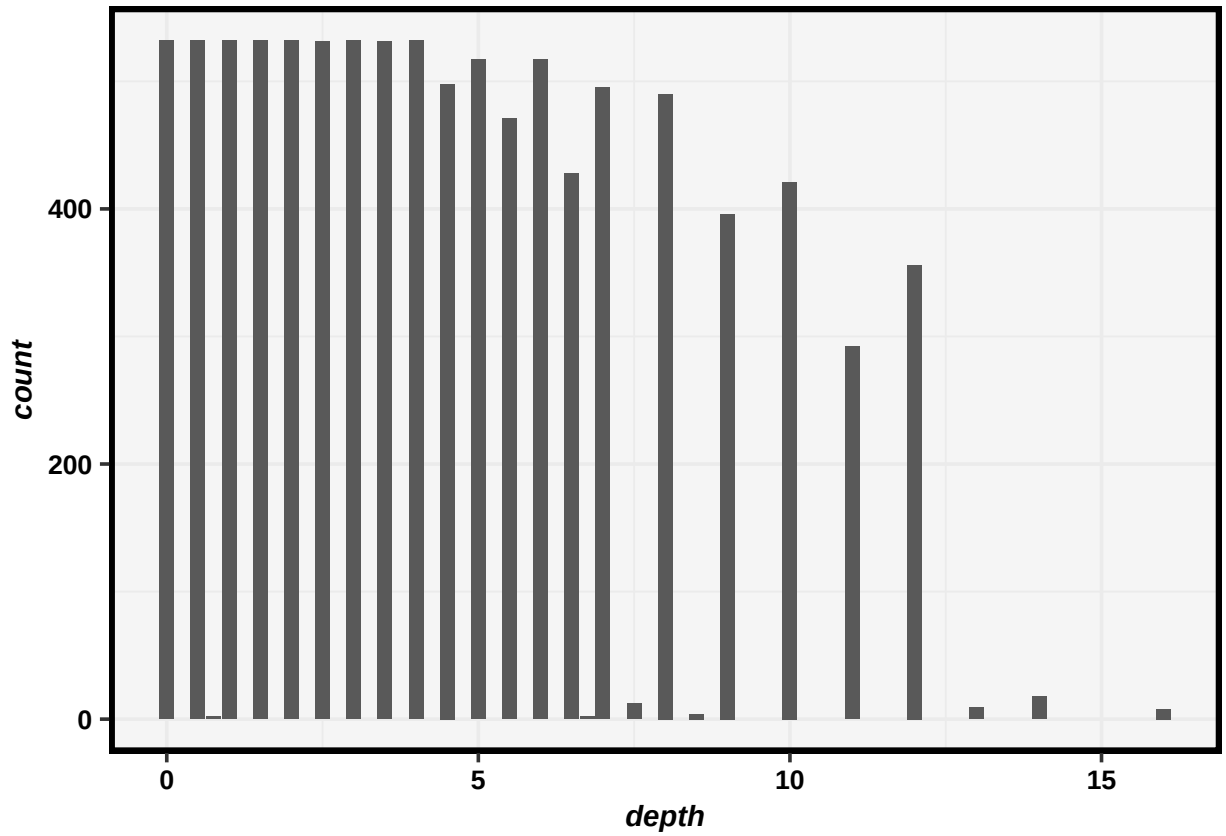
```
ggplot(PeterPaulJulyFiltered, aes(x = year4)) +
  geom_histogram(stat = "count")
```




```
ggplot(PeterPaulJulyFiltered, aes(x =daynum)) +
  geom_histogram(stat = "count")
```




```
ggplot(PeterPaulJulyFiltered, aes(x =depth)) +
  geom_histogram(stat = "count")
```




## 12

Run an AIC to determine what set of explanatory variables (year4, daynum, depth) is best suited to predict temperature.

```
TempMod<-lm(data=PeterPaulJulyFiltered, temperature_C~year4+daynum+depth)

summary(TempMod)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = PeterPaulJulyFiltered)
##
## 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
```

## Use step function

```
step(TempMod)
```

```
## 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 = PeterPaulJulyFiltered)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##    -6.45556      0.01013      0.04134     -1.94726
```

**Fullest model is best according to this Stepwise Algorithm function**

## Check using Stepwise Model Reduction Process

### Remove year4

```
TempMod2<-update(TempMod,.~.-year4)
```

### Summary

```
summary(TempMod2)
```

```
##
## Call:
## lm(formula = temperature_C ~ daynum + depth, data = PeterPaulJulyFiltered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6197 -2.9772  0.0797  2.9616 13.4633
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 13.782641   0.854151   16.136   <2e-16 ***
## daynum       0.041389   0.004316    9.591   <2e-16 ***
## depth       -1.946955   0.011678 -166.727   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.812 on 9719 degrees of freedom
## Multiple R-squared:  0.7416, Adjusted R-squared:  0.7415
## F-statistic: 1.395e+04 on 2 and 9719 DF,  p-value: < 2.2e-16
```

### Test AIC Scores

```
AIC(TempMod, TempMod2)
```

```
##          df      AIC
## TempMod   5 53608.15
## TempMod2  4 53611.70
```

#### Fullest Model is best

### Run a multiple regression on the recommended set of variables.

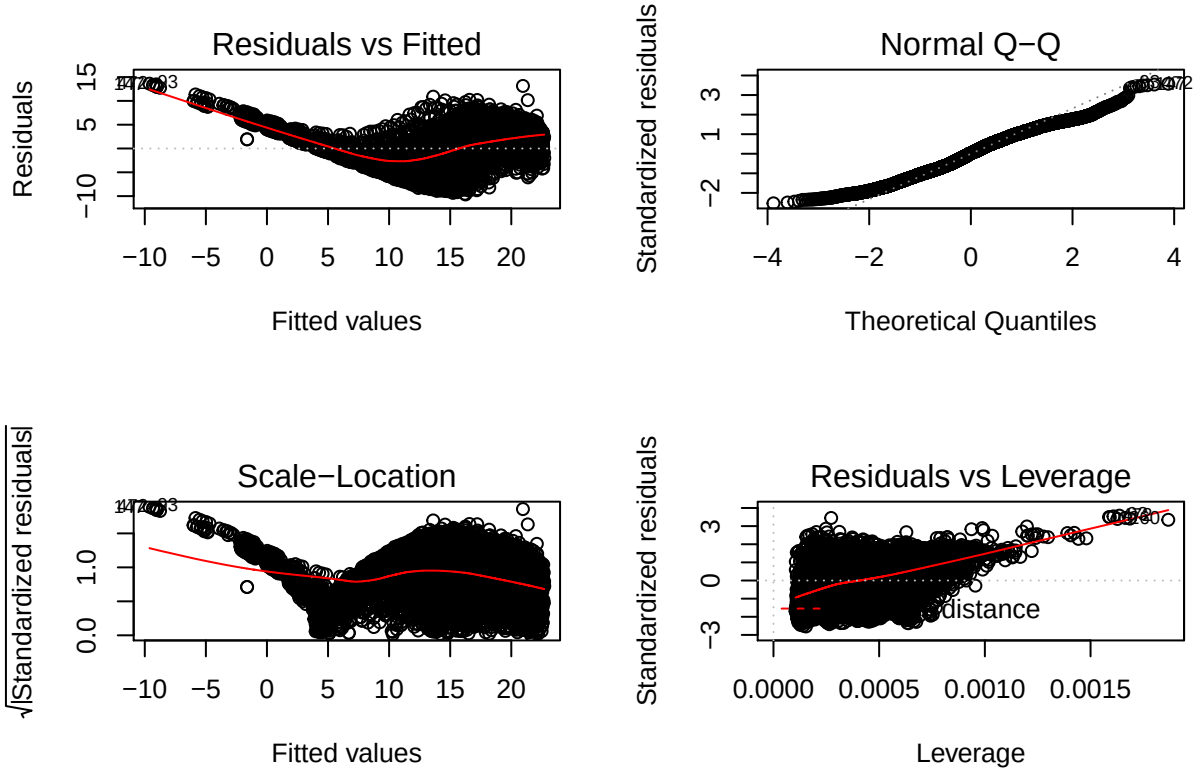
```
TempMod<-lm(data=PeterPaulJulyFiltered, temperature_C~year4+daynum+depth)

summary(TempMod)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = PeterPaulJulyFiltered)
##
## 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
```

#### Check Residuals of Final Model

```
par(mfrow=c(2,2))
plot(TempMod)
```




## 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:
> Temperature=-6.45+0.01(year) +0.04(daynum)-1.95(depth) + E. The Rsquared value is 0.74, which means my final model accounts for 74% of the variance of the dependent variable, temperature. (Test: Multiple Regression, p<2.2e-16, df=9718, R=0.74).

## 14

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

```
TempANCOVAMod<-lm(temperature_C~depth*lakename, data=PeterPaulJulyFiltered)
summary(TempANCOVAMod)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth * lakename, data = PeterPaulJulyFiltered)
##
## 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**

```
summary.aov(TempANCOVAMod)
```

```
##                   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: If the p-value for the interaction effect is less than 0.05, then we would consider the interaction among the explanatory variables to be significant. According to the above ANOVA summary using summary.aov function, there is a significant interaction bectween depth and lakename because the p value<2e-16 (Linear Regression; p<2e-16, df=8, F=48.64). Therefore, the interaction between depth and lake name is significant and should be included in the model. The adjusted r-squared, the modified R-sqaured adjusted for the number of predictors in the

model, is 0.79, and thus, 79% of the variance in the response variable can be explained by the predictor variables.

## 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.

```
TempbyDepthPlot<-ggplot(PeterPaulJulyFiltered, aes(x =depth, y =temperature_C, color=lakename)) +
  geom_point(aes(x=depth, y=temperature_C), size=0.7, shape=16, alpha=0.5) +
    geom_smooth(aes(x =depth, y =temperature_C, span=0.1, color=lakename),
                method="lm", se=FALSE, linetype=1, size=0.5) +
  labs(title="The Effect of Depth on Temperature", x="Depth of Lake (meters)",
       y="Temperature (degrees C)") + gabytheme+

  theme(axis.text.y=element_text(angle = 35,  hjust = 1)) +
 scale_color_brewer(palette = "Paired")
  print(TempbyDepthPlot)
```