Assignment 6: Generalized Linear Models

Keith Bollt

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.

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

```
#1
getwd()
```

```
## [1] "V:/ENV_872_Project_Directory/Assignments"
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.2
```

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

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     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
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
## Warning: package 'tibble' was built under R version 3.5.2
```

```
## Warning: package 'tidyr' was built under R version 3.5.2
```

```
## Warning: package 'readr' was built under R version 3.5.2
```

```
## Warning: package 'purrr' was built under R version 3.5.2
```

```
## Warning: package 'dplyr' was built under R version 3.5.2
```

```
## Warning: package 'stringr' was built under R version 3.5.2
```

```
## Warning: package 'forcats' was built under R version 3.5.2
```

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

```
library(ggplot2)
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.5.2
```

```
## corrplot 0.84 loaded
```

```
knitr::opts_chunk$set(fig.height = 9, fig.width = 7)

Neonics.A06 <-
  read.csv("V:/ENV_872_Project_Directory/Data/Raw/ECOTOX_Neonicotinoids_Mortality_raw.csv")
chemistry.physics.raw.A06 <-
  read.csv("V:/ENV_872_Project_Directory/Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv")

#2
mytheme.A06 <-
  theme_classic(base_size = 14)+
  theme(axis.text = element_text(color = "blue"),
  legend.position = "top")
theme_set(mytheme.A06)
```

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

```
#3
Neonics.A06$Chemical.Name <- as.character(Neonics.A06$Chemical.Name)

choices <- length(unique(Neonics.A06$Chemical.Name))
summary(choices)  # The answer is 9 different chemicals
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       9       9       9       9       9       9
```

```
#4
summary(Neonics.A06$Chemical.Name) # This generates the different chemical names. I should have just ru
```

```
##    Length     Class      Mode
##      1283 character character
```

```r
shapiro.test(Neonics.A06$Pub..Year[Neonics.A06$Chemical.Name == "Acetamiprid"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Neonics.A06$Pub..Year[Neonics.A06$Chemical.Name == "Acetamiprid"]
## W = 0.90191, p-value = 5.706e-08
```

```r
shapiro.test(Neonics.A06$Pub..Year[Neonics.A06$Chemical.Name == "Clothianidin"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Neonics.A06$Pub..Year[Neonics.A06$Chemical.Name == "Clothianidin"]
## W = 0.69577, p-value = 4.287e-11
```

```r
shapiro.test(Neonics.A06$Pub..Year[Neonics.A06$Chemical.Name == "Dinotefuran"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Neonics.A06$Pub..Year[Neonics.A06$Chemical.Name == "Dinotefuran"]
## W = 0.82848, p-value = 8.83e-07
```

```r
shapiro.test(Neonics.A06$Pub..Year[Neonics.A06$Chemical.Name == "Imidacloprid"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Neonics.A06$Pub..Year[Neonics.A06$Chemical.Name == "Imidacloprid"]
## W = 0.88178, p-value < 2.2e-16
```

```r
shapiro.test(Neonics.A06$Pub..Year[Neonics.A06$Chemical.Name == "Imidaclothiz"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Neonics.A06$Pub..Year[Neonics.A06$Chemical.Name == "Imidaclothiz"]
## W = 0.68429, p-value = 0.00093
```

```r
shapiro.test(Neonics.A06$Pub..Year[Neonics.A06$Chemical.Name == "Nitenpyram"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Neonics.A06$Pub..Year[Neonics.A06$Chemical.Name == "Nitenpyram"]
## W = 0.79592, p-value = 0.0005686
```

```r
shapiro.test(Neonics.A06$Pub..Year[Neonics.A06$Chemical.Name == "Nithiazine"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Neonics.A06$Pub..Year[Neonics.A06$Chemical.Name == "Nithiazine"]
## W = 0.75938, p-value = 0.0001235
```

```r
shapiro.test(Neonics.A06$Pub..Year[Neonics.A06$Chemical.Name == "Thiacloprid"])
```

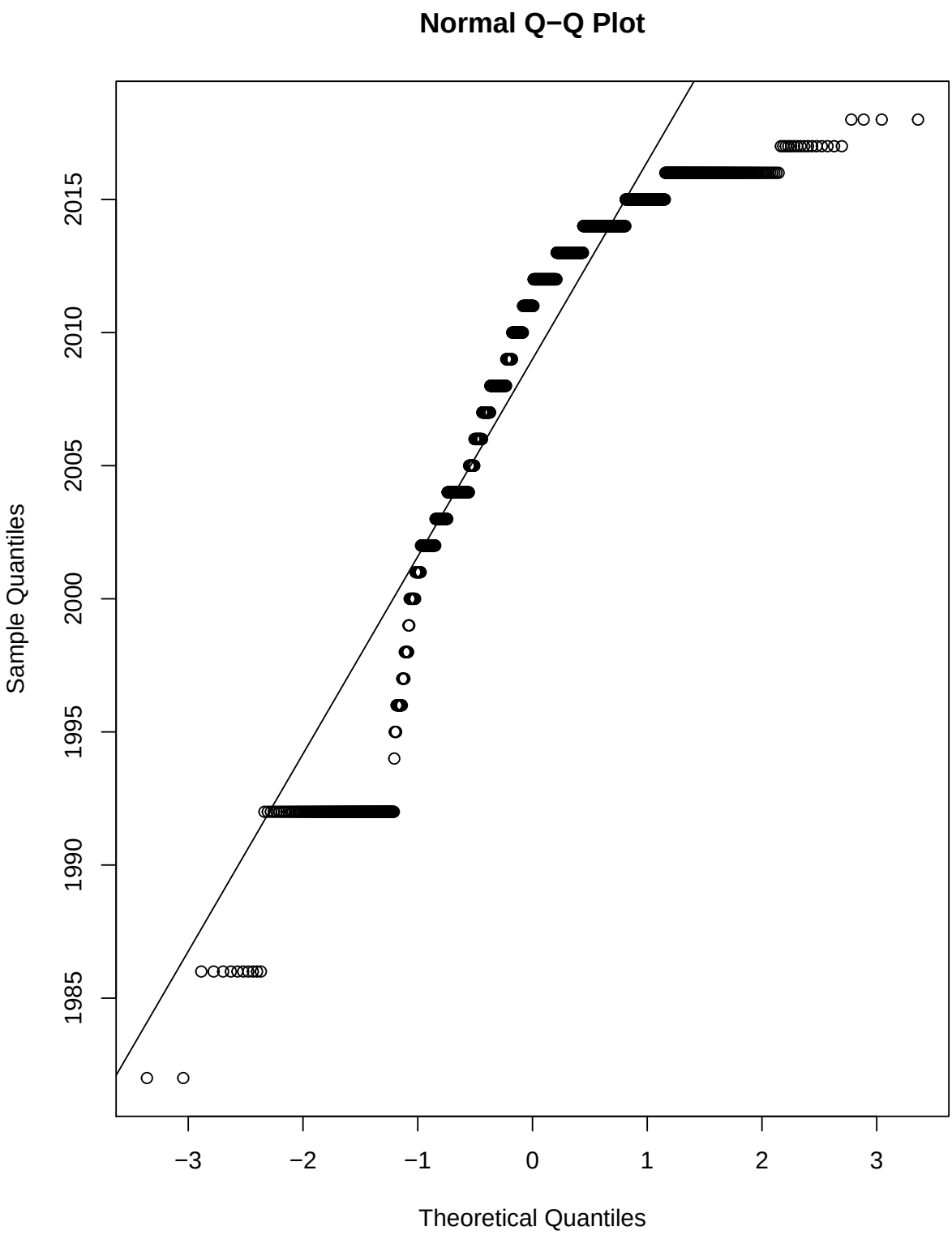
```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  Neonics.A06$Pub..Year[Neonics.A06$Chemical.Name == "Thiacloprid"]
## W = 0.7669, p-value = 1.118e-11
```

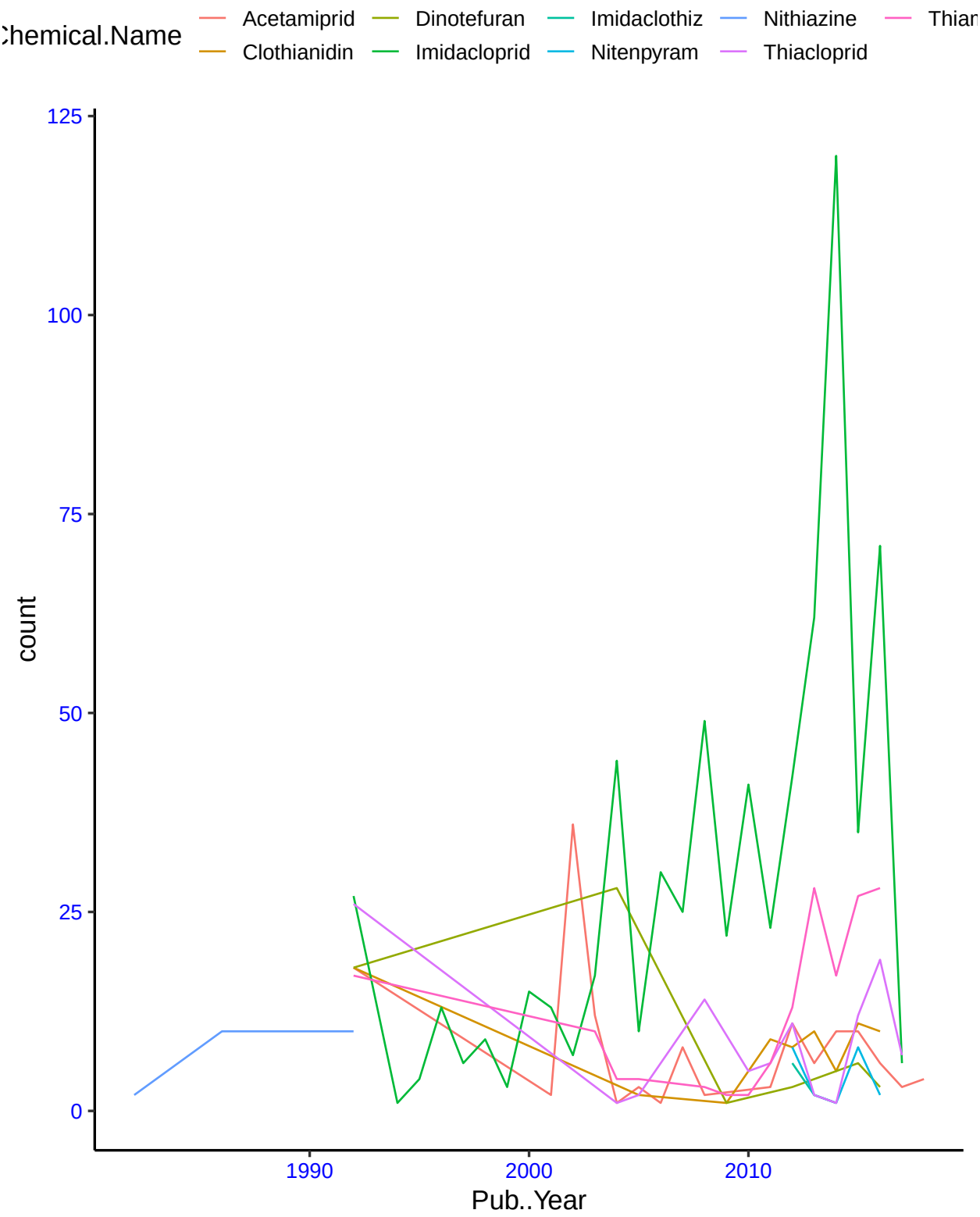
```
shapiro.test(Neonics.A06$Pub..Year[Neonics.A06$Chemical.Name == "Thiamethoxam"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Neonics.A06$Pub..Year[Neonics.A06$Chemical.Name == "Thiamethoxam"]
## W = 0.7071, p-value < 2.2e-16
```

```
q4.plot <- ggplot(Neonics.A06, aes(x = Pub..Year, color = Chemical.Name)) +
  geom_freqpoly(stat = "count")
qqnorm(Neonics.A06$Pub..Year); qqline(Neonics.A06$Pub..Year)
```




```
print(q4.plot)
```




```
#No. The p-value for each Shapiro Test is less than 0.05, so we reject the null hypothesis that the dat

#5
```

```
neonics.kw <- kruskal.test(Neonics.A06$Chemical.Name ~ Neonics.A06$Pub..Year)
neonics.kw
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Neonics.A06$Chemical.Name by Neonics.A06$Pub..Year
## Kruskal-Wallis chi-squared = 164.61, df = 27, p-value < 2.2e-16
```

```
# No. The p-value < 0.05, so I reject the null hypothesis that there is no significant difference betwe
```

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

   ANSWER: Dunn Test

7. Run this test below.

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

```
#7
library(FSA)
```

```
## Warning: package 'FSA' was built under R version 3.5.2
```

```
## ## FSA v0.8.22. See citation('FSA') if used in publication.
## ## Run fishR() for related website and fishR('IFAR') for related book.
```

```
dunnTest(Neonics.A06$Pub..Year, Neonics.A06$Chemical.Name)
```

```
## Warning: 'g' variable was coerced to a factor.
```

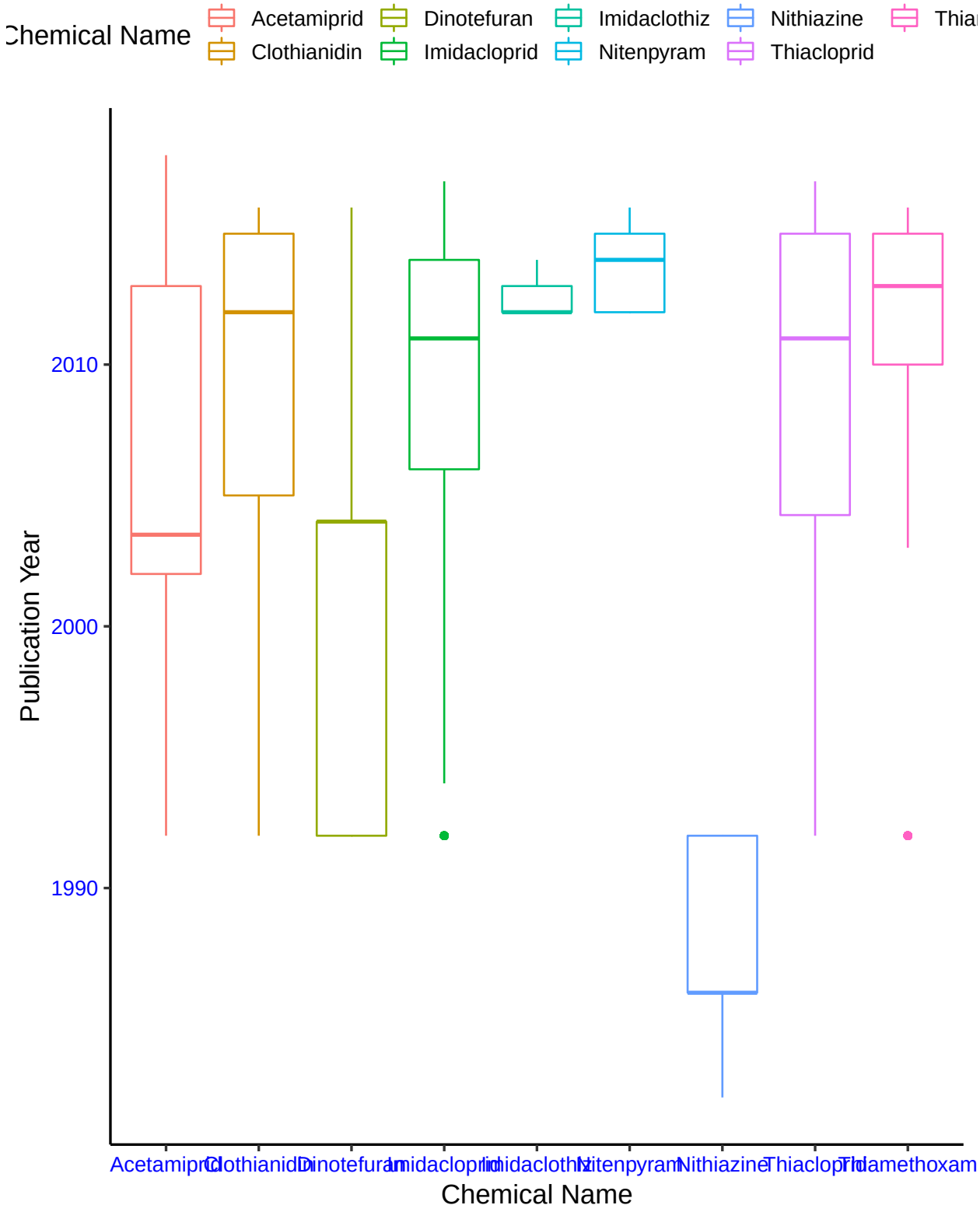
```
## Dunn (1964) Kruskal-Wallis multiple comparison
```

```
##   p-values adjusted with the Holm method.
```

```
##                      Comparison          Z      P.unadj        P.adj
## 1   Acetamiprid - Clothianidin -3.0388079 2.375163e-03 4.037777e-02
## 2    Acetamiprid - Dinotefuran  2.1172089 3.424212e-02 4.109054e-01
## 3   Clothianidin - Dinotefuran  4.4060765 1.052598e-05 2.420975e-04
## 4   Acetamiprid - Imidacloprid -4.0204987 5.807507e-05 1.277651e-03
## 5  Clothianidin - Imidacloprid  0.5068899 6.122321e-01 1.000000e+00
## 6   Dinotefuran - Imidacloprid -5.2140290 1.847826e-07 4.989129e-06
## 7   Acetamiprid - Imidaclothiz -1.8052932 7.102881e-02 7.813169e-01
## 8  Clothianidin - Imidaclothiz -0.5166649 6.053901e-01 1.000000e+00
## 9   Dinotefuran - Imidaclothiz -2.6586494 7.845456e-03 1.176818e-01
## 10 Imidacloprid - Imidaclothiz -0.7284284 4.663514e-01 1.000000e+00
## 11    Acetamiprid - Nitenpyram -4.5018639 6.736012e-06 1.616643e-04
## 12   Clothianidin - Nitenpyram -2.4936264 1.264456e-02 1.770238e-01
## 13    Dinotefuran - Nitenpyram -5.4527796 4.958852e-08 1.388479e-06
## 14   Imidacloprid - Nitenpyram -3.0634837 2.187761e-03 3.937970e-02
## 15   Imidaclothiz - Nitenpyram -1.0897204 2.758363e-01 1.000000e+00
## 16    Acetamiprid - Nithiazine  5.6425299 1.675694e-08 4.859513e-07
## 17   Clothianidin - Nithiazine  7.1473251 8.848514e-13 2.831524e-11
## 18    Dinotefuran - Nithiazine  3.8693508 1.091255e-04 2.291636e-03
## 19   Imidacloprid - Nithiazine  7.7286349 1.087060e-14 3.804708e-13
## 20   Imidaclothiz - Nithiazine  4.8473136 1.251445e-06 3.253758e-05
## 21     Nitenpyram - Nithiazine  7.7099812 1.258363e-14 4.278434e-13
## 22   Acetamiprid - Thiacloprid -3.2225618 1.270497e-03 2.413945e-02
## 23  Clothianidin - Thiacloprid  0.1414916 8.874816e-01 8.874816e-01
```

```
## 24   Dinotefuran - Thiacloprid -4.6025295 4.173904e-06 1.043476e-04
## 25  Imidacloprid - Thiacloprid -0.3888712 6.973714e-01 1.000000e+00
## 26  Imidaclothiz - Thiacloprid  0.5870686 5.571576e-01 1.000000e+00
## 27    Nitenpyram - Thiacloprid  2.6709745 7.563140e-03 1.210102e-01
## 28    Nithiazine - Thiacloprid -7.3166886 2.541647e-13 8.387437e-12
## 29  Acetamiprid - Thiamethoxam -5.8898861 3.864618e-09 1.159385e-07
## 30 Clothianidin - Thiamethoxam -1.7587256 7.862413e-02 7.862413e-01
## 31  Dinotefuran - Thiamethoxam -6.6762123 2.451967e-11 7.601098e-10
## 32 Imidacloprid - Thiamethoxam -3.5327039 4.113329e-04 8.226657e-03
## 33 Imidaclothiz - Thiamethoxam -0.1886278 8.503846e-01 1.000000e+00
## 34   Nitenpyram - Thiamethoxam  1.5927766 1.112103e-01 1.000000e+00
## 35   Nithiazine - Thiamethoxam -8.7224129 2.723352e-18 9.804067e-17
## 36  Thiacloprid - Thiamethoxam -2.1461156 3.186376e-02 4.142288e-01
```

```r
#8
Neonics.boxplot <- ggplot(Neonics.A06, aes(x = Chemical.Name, y = Pub..Year, color = Chemical.Name) )+
  geom_boxplot()+
  xlab(expression("Chemical Name"))+
  ylab(expression("Publication Year"))+
  labs(color = "Chemical Name")
print(Neonics.boxplot)
```



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: The studies on the various neonicotides were conducted in different years. (After running a Dunn test on the nonnormal data distribution of chemical vs publication year, I found that the relationship between every chemical pairing combination had significantly different publication years. The z-scores were both positive and negative depending on the pairings, but the most important finding from the Dunn test was that the p-values were below 0.05 for every combination.)

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)

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.

```
#11
chemistry.physics.processed.A06 <-
  chemistry.physics.raw.A06 %>%
  filter(daynum == 182:212) %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  na.omit()
```

```
## Warning: package 'bindrcpp' was built under R version 3.5.2
```

```
## Warning in daynum == 182:212: longer object length is not a multiple of
## shorter object length
```

```
#12
chemistry.A06.AIC <- lm(data = chemistry.physics.processed.A06, temperature_C ~ year4 + daynum +
              depth)
step(chemistry.A06.AIC)
```

```
## Start:  AIC=845.18
## temperature_C ~ year4 + daynum + depth
##
##          Df Sum of Sq     RSS     AIC
## - year4   1      12.0  4577.5  844.00
## - daynum  1      28.2  4593.7  845.10
## <none>                 4565.5  845.18
## - depth   1   12873.2 17438.6 1261.31
##
## Step:  AIC=844
## temperature_C ~ daynum + depth
##
##          Df Sum of Sq     RSS     AIC
## - daynum  1      27.5  4605.0  843.87
## <none>                 4577.5  844.00
## - depth   1   12863.0 17440.5 1259.35
##
## Step:  AIC=843.87
## temperature_C ~ depth
##
```

```
##         Df Sum of Sq   RSS     AIC
## <none>                4605  843.87
## - depth  1     12848 17453 1257.57
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = chemistry.physics.processed.A06)
##
## Coefficients:
## (Intercept)        depth
##      22.008       -1.961
```

```
# the lower the aic, the better the correlation. Depth by itself produced the lowest AIC, and therefore

temperature.model <- lm(data = chemistry.physics.processed.A06, temperature_C ~ depth)
summary(temperature.model)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = chemistry.physics.processed.A06)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8450 -3.0840  0.1782  3.0398 13.3673
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.00846    0.38752   56.79   <2e-16 ***
## depth       -1.96099    0.06668  -29.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.854 on 310 degrees of freedom
## Multiple R-squared:  0.7362, Adjusted R-squared:  0.7353
## F-statistic: 864.9 on 1 and 310 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: temperature_C = 22 - (1.96 * depth) + error This equation has an r^2 value of about 0.73, so this equation explains about 73% of the observed variance.

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

```
#14
temperature.ancova.interaction <- lm(data = chemistry.physics.processed.A06, temperature_C ~ depth * lak
summary(temperature.ancova.interaction)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth * lakename, data = chemistry.physics.processed.A06)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6100 -2.7826 -0.2609  2.8225 12.1803
##
## Coefficients: (1 not defined because of singularities)
```

```
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      19.54590    2.66139   7.344 2.03e-12 ***
## depth                            -1.83647    0.19244  -9.543  < 2e-16 ***
## lakenameCrampton Lake             4.35631    3.13662   1.389   0.1659
## lakenameEast Long Lake           -1.26077    2.87928  -0.438   0.6618
## lakenameHummingbird Lake         -0.04503    3.81149  -0.012   0.9906
## lakenamePaul Lake                 3.76866    2.76530   1.363   0.1740
## lakenamePeter Lake                3.98209    2.73251   1.457   0.1461
## lakenameTuesday Lake              0.59795    2.84690   0.210   0.8338
## lakenameWard Lake                 8.38017    5.40085   1.552   0.1218
## lakenameWest Long Lake            1.12336    2.62700   0.428   0.6692
## depth:lakenameCrampton Lake      -0.04629    0.36216  -0.128   0.8984
## depth:lakenameEast Long Lake      0.22998    0.27078   0.849   0.3964
## depth:lakenameHummingbird Lake -0.64710    0.64119  -1.009   0.3137
## depth:lakenamePaul Lake          -0.31070    0.23393  -1.328   0.1851
## depth:lakenamePeter Lake         -0.14529    0.21836  -0.665   0.5063
## depth:lakenameTuesday Lake       -0.07844    0.25389  -0.309   0.7576
## depth:lakenameWard Lake          -1.91234    1.01063  -1.892   0.0594 .
## depth:lakenameWest Long Lake           NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.603 on 295 degrees of freedom
## Multiple R-squared:  0.7806, Adjusted R-squared:  0.7687
## F-statistic: 65.59 on 16 and 295 DF,  p-value: < 2.2e-16
```

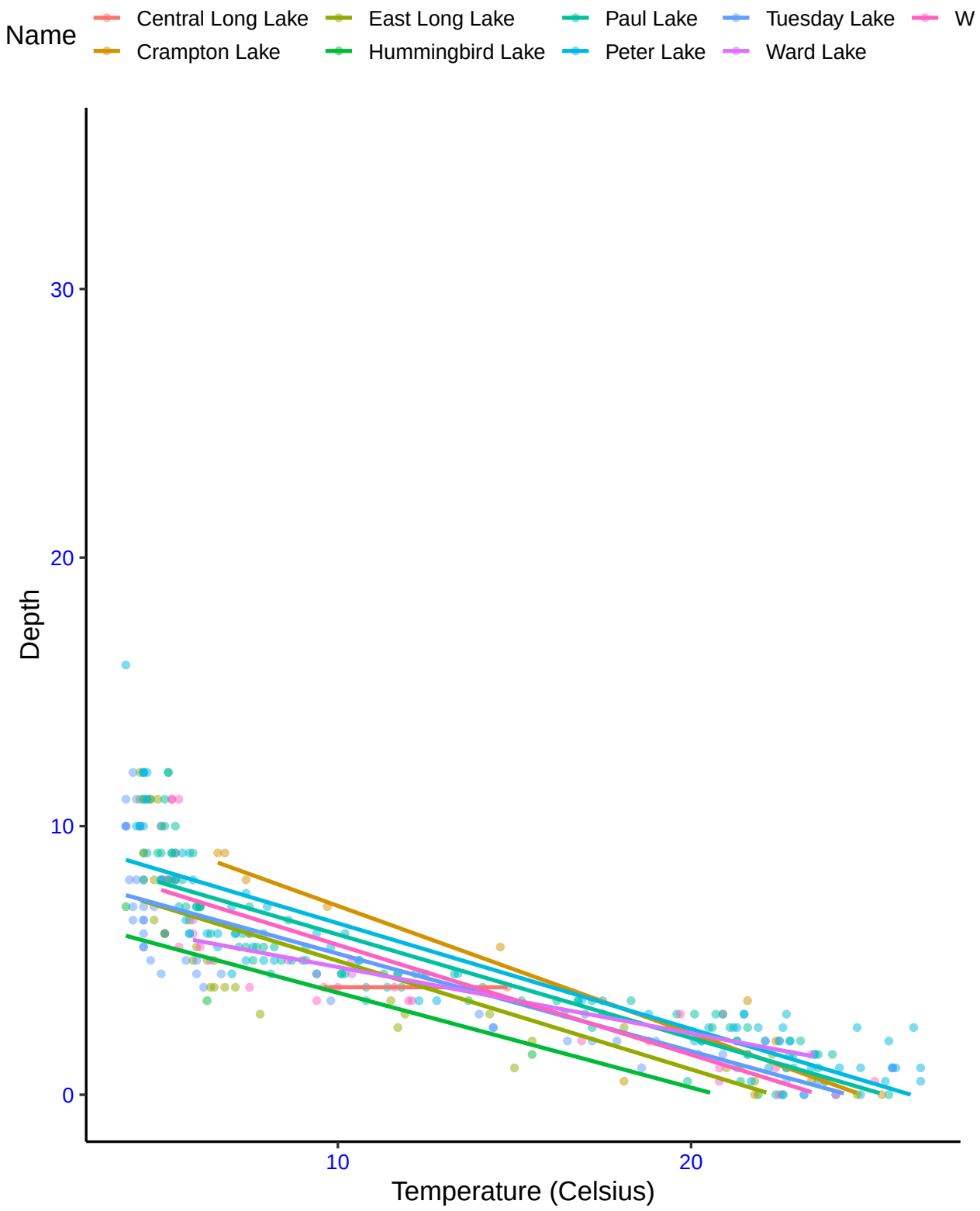
15. Is there an interaction between depth and lakename? How much variance in the temperature observations does this explain?

    ANSWER: There is not a significant interaction between depth and lakename. The depth:lakename interaction for every lake in the dataset had a Pr(>|t|) value (basically a p-value) greater than 0.05. This model explains 77 percent of the variance in the 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
temperature.vs.depth.plot <- ggplot(chemistry.physics.processed.A06, aes(x = temperature_C, y = depth,
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  ylim(0, 35)+
  labs(x= "Temperature (Celsius)", y = "Depth", color = "Lake Name")
print(temperature.vs.depth.plot)
```

```
## Warning: Removed 33 rows containing missing values (geom_smooth).
```


Name
Central Long Lake
Crampton Lake
East Long Lake
Hummingbird Lake
Paul Lake
Peter Lake
Tuesday Lake
Ward Lake
W
Depth
30
20
10
0
10
20
Temperature (Celsius)