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

Austin Guimond

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
knitr::opts_chunk$set(tidy.opts=list(width.cutoff=80), tidy=TRUE)
getwd()
```

```
## [1] "/home/guest/R/EDA-Fall2022"
```

```
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   1.2.0      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(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library (cowplot)
```

```
##
## Attaching package: 'cowplot'
##
## The following object is masked from 'package:lubridate':
##
##     stamp
```

```
#import lake data
LakeChem <-
read.csv("~/R/EDA-Fall2022/Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv", stringsAsFactors = TRUE)
LakeChem$sampledate <- as.Date(LakeChem$sampledate, format = "%m/%d/%y")

#2
#set theme
mytheme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "bottom")
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: Mean lake temperature does not chnage with depth across all lakes in July. Ha: Mean lake temperatures vary with depth accross lakes in July.
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.

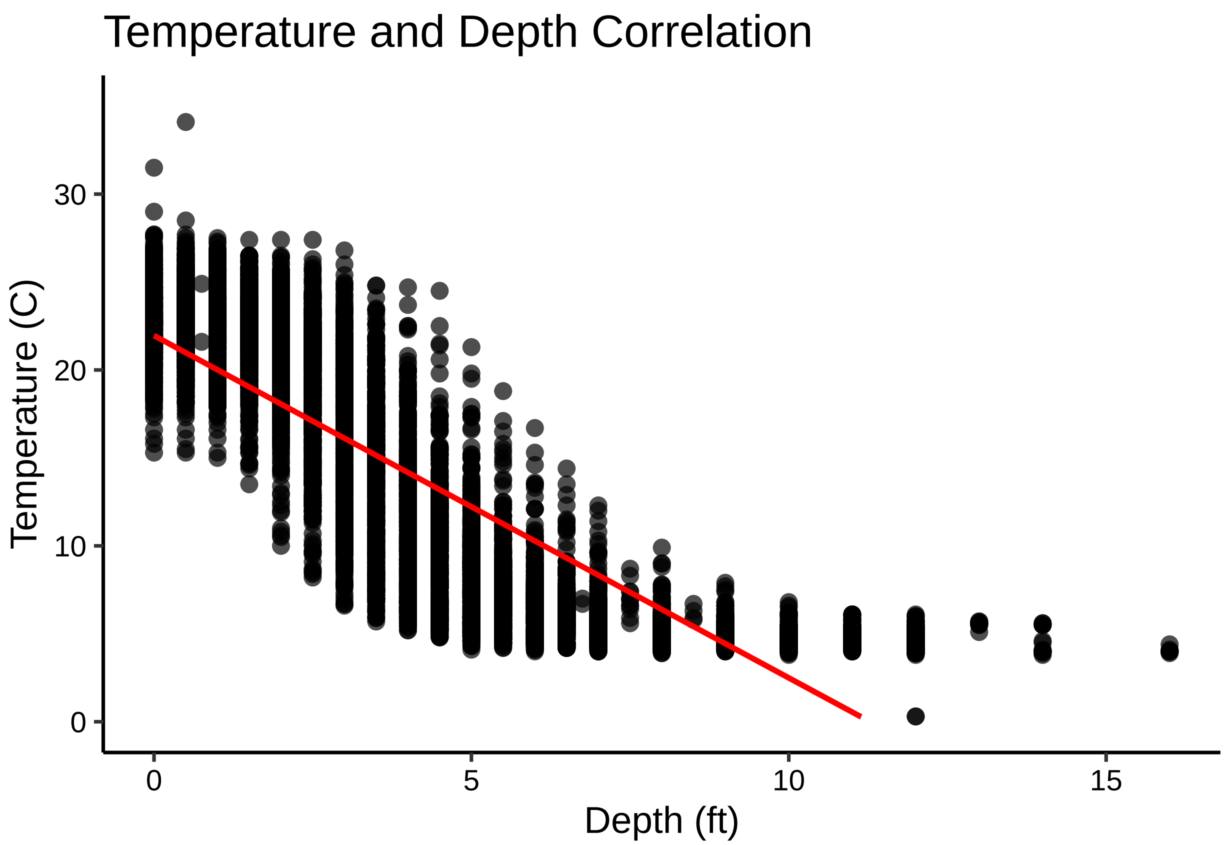
```
# 4 wrangle data for July with only year, day, depth and temperature
LakeChem.Pipe <- mutate(LakeChem, month = month(sampledate)) %>%
    filter(month == 7) %>%
    select(lakename, year4, daynum, depth, temperature_C) %>%
```

```
    na.omit()

# 5 create chart shwoing correlation of temperature and depth
Depth.Temp.Chart <- ggplot(LakeChem.Pipe, aes(x = depth, y = temperature_C)) + geom_point(alpha = 0.7,
    size = 2.5) + ylim(0, 35) + geom_smooth(method = lm, col = "red") + xlab("Depth (ft)") +
    ylab("Temperature (C)") + ggtitle("Temperature and Depth Correlation")
print(Depth.Temp.Chart)
```

```
## `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 indicates that as depth increases the lake temperature decreases. This is further displayed by the trendline that has a downward slope. The cluster of data is mostly between 0 and 8 feet deep. This indiccates that there are more data points at shallower depths. The clustering of the points could also indcate that the decrease in temperature is not a linear trend. The point distribution seems to indciate that the decrease in temperature is linear until 10 feet where the tempertur stays around 7-5 degrees C despite changes in depth.

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

```r
# 7 temperature and depth linear regression
depth.regression <- lm(LakeChem.Pipe$temperature_C ~ LakeChem.Pipe$depth)
summary(depth.regression)
```

```
##
## Call:
## lm(formula = LakeChem.Pipe$temperature_C ~ LakeChem.Pipe$depth)
##
## 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 ***
## LakeChem.Pipe$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 R squared value for the regression is 0.7387, indicating that depth accounts for 74% of the variability in temperature. The degrees of freedom is 9727 which is based on the the number of observations that were used in the analysis. The P value of for the test is well below .05 indicating that the correlation between temperture and depth is significant. The slope of the regression line is -1.95. Using a graph, we could track how the temperature would change in the lake by depth. The slope value indicates that for each drop in 1m of depth, the temperature goes down by 1.95 degrees C.

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.

```r
# 9
AIC <- lm(data = LakeChem.Pipe, temperature_C ~ year4 + daynum + depth)

step(AIC)
```

```
## 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 = LakeChem.Pipe)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##    -8.57556      0.01134      0.03978     -1.94644
```

```
AIC_Model <- lm(data = LakeChem.Pipe, temperature_C ~ year4 + daynum + depth)
summary(AIC_Model)
```

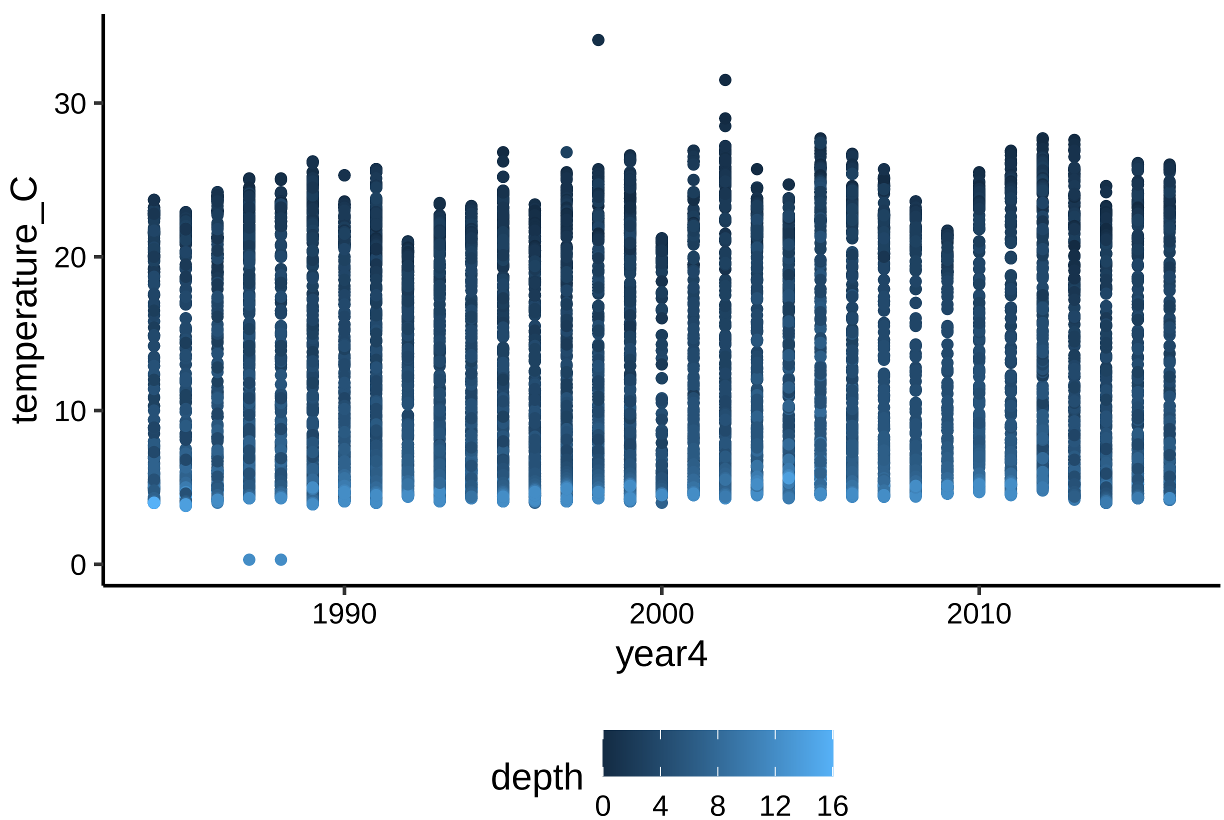
```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = LakeChem.Pipe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6536 -3.0000  0.0902  2.9658 13.6123
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -8.575564   8.630715   -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
```

```
# 10 Multiple regression with year, day, and depth.  All values had a
# statistically significant P value for impacting lake temperature
MultiRegression <- lm(data = LakeChem.Pipe, temperature_C ~ year4 + daynum + depth)
summary(MultiRegression)
```

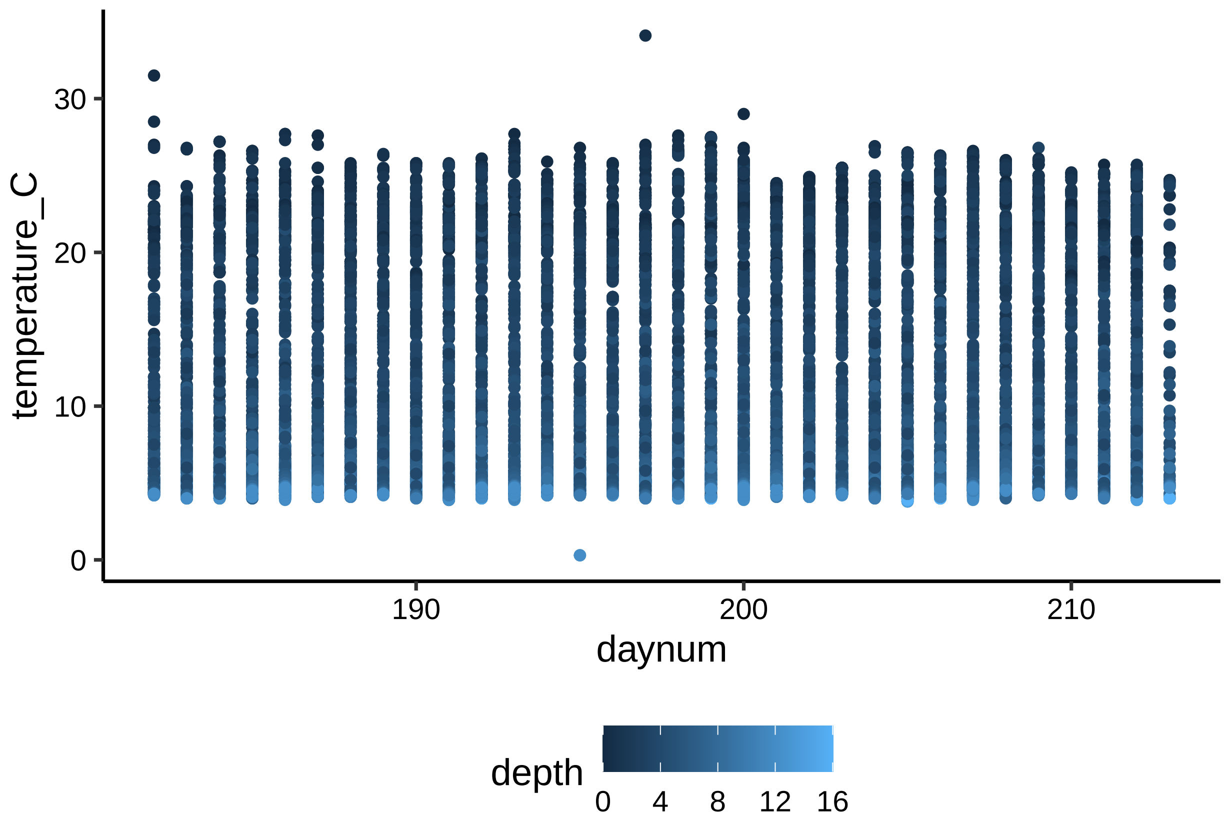
```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = LakeChem.Pipe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -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
```

```
MutliRegressionYear <- ggplot(LakeChem.Pipe) + aes(x = year4, y = temperature_C,
    color = depth) + geom_point()
print(MutliRegressionYear)
```




```
MutliRegressionDay <- ggplot(LakeChem.Pipe) + aes(x = daynum, y = temperature_C,
    color = depth) + geom_point()
print(MutliRegressionDay)
```



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 suggests that year, day, and depth should all be used to predict changes in lake temperature. All the variables have P values lower than .05 indicating that there is a significant correlation. However, the P value for year was higher compared to the other two variables. This indicates that changes in day and depth are most closely tied to changes in lake temperature. The model with all three variables explains 74 perecent of the variation in temperature based on the r squared value of 0.7412. This is a very slight improvement compared to the r squared value for just depth but not by much. Although day and year are significantly correlated to temperature, most of the temperature variation stems from changes in depth.

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).

```r
# 12
library(agricolae)
# group data by lake name
Lake.Temps <- LakeChem.Pipe %>%
    group_by(lakename) %>%
    summarise(temperature_C)
```

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

```r
summary(Lake.Temps)
```

```
##             lakename    temperature_C
##  Peter Lake    :2872   Min.   : 0.30
##  Paul Lake     :2660   1st Qu.: 5.50
##  Tuesday Lake  :1524   Median :10.10
##  West Long Lake:1026   Mean   :12.72
##  East Long Lake: 968   3rd Qu.:20.80
##  Crampton Lake : 318   Max.   :34.10
##  (Other)       : 360
```

```r
summary(Lake.Temps$lakename)
```

```
## Central Long Lake     Crampton Lake    East Long Lake  Hummingbird Lake
##               128               318               968               116
##         Paul Lake        Peter Lake      Tuesday Lake         Ward Lake
##              2660              2872              1524               116
##    West Long Lake
##              1026
```

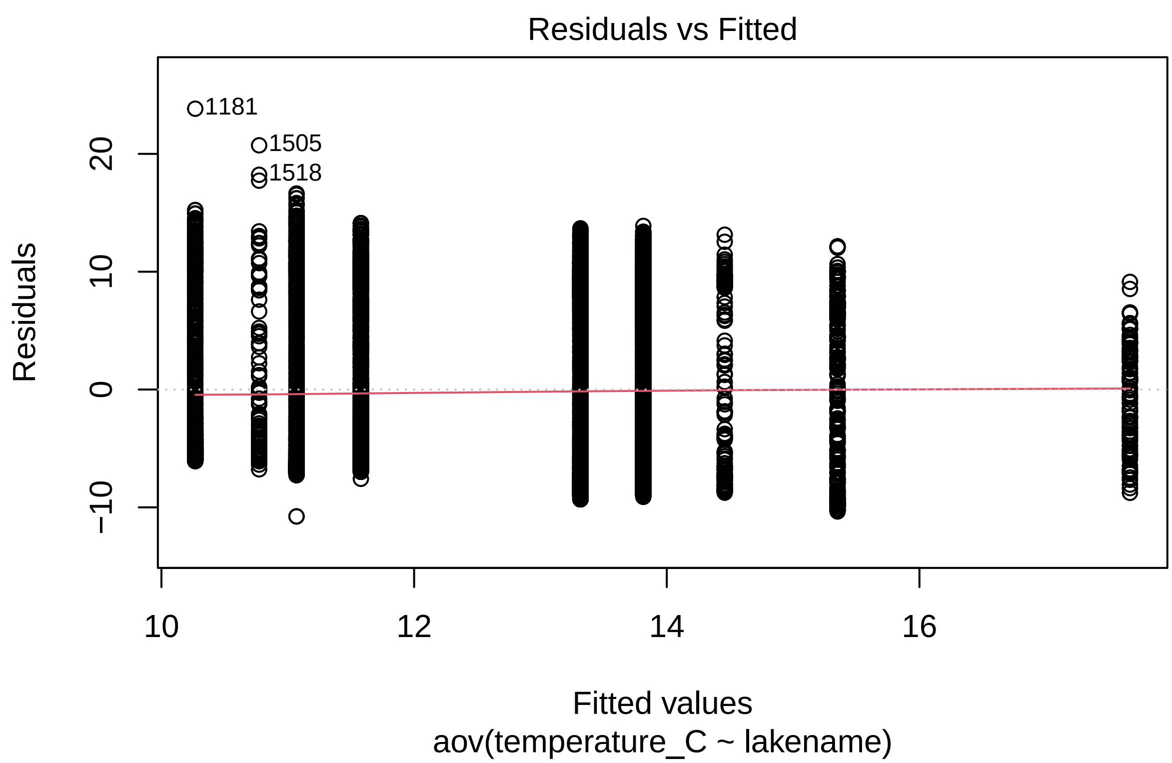
```r
bartlett.test(Lake.Temps$temperature_C ~ Lake.Temps$lakename)
```

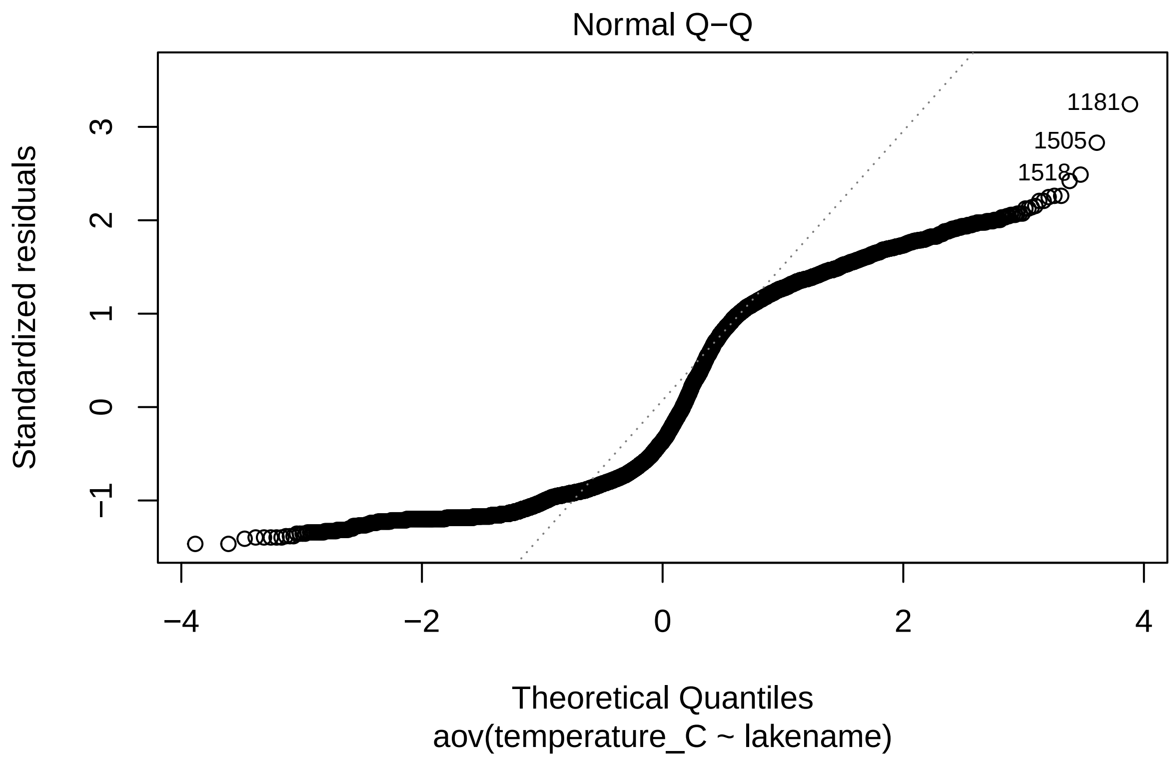
```
##
##  Bartlett test of homogeneity of variances
##
## data:  Lake.Temps$temperature_C by Lake.Temps$lakename
## Bartlett's K-squared = 92.777, df = 8, p-value < 2.2e-16
```

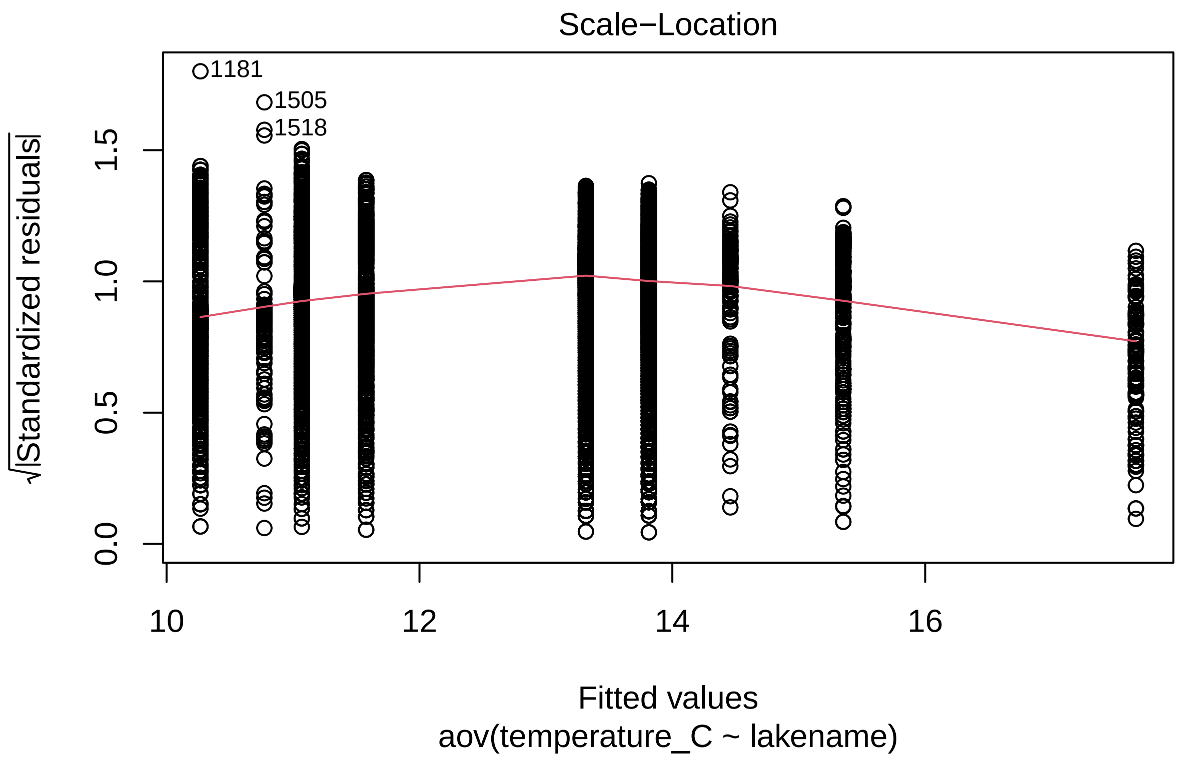
```r
# anova test for temperature and lake
Lake.Temp_Anova <- aov(data = Lake.Temps, temperature_C ~ lakename)
summary(Lake.Temp_Anova)
```

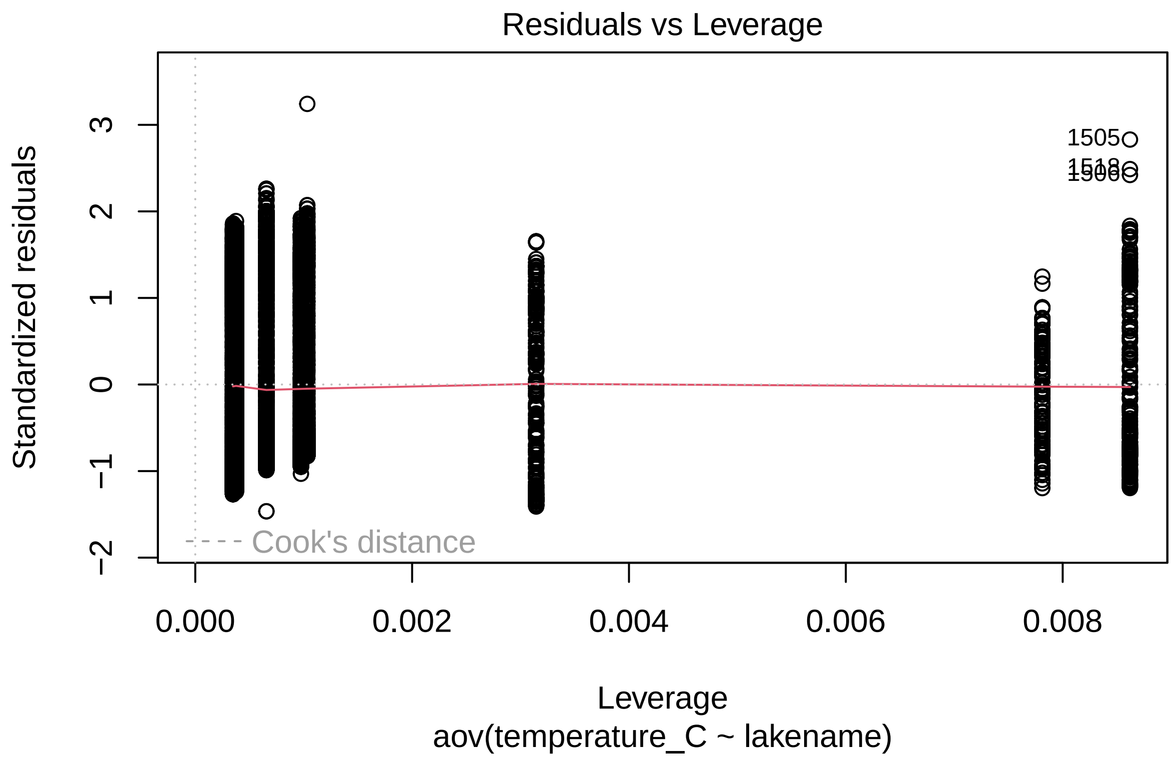
```
##               Df Sum Sq Mean Sq F value Pr(>F)
## lakename       8  21642  2705.2      50 <2e-16 ***
## Residuals   9719 525813    54.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(Lake.Temp_Anova)
```




Normal Q−Q
1181
1505
1518
3
2
1
0
−1
Standardized residuals
−4
−2
0
2
4
Theoretical Quantiles
aov(temperature_C ~ lakename)


Scale−Location
1181
1505
1518
1.5
1.0
0.5
0.0
√|Standardized residuals|
10
12
14
16
Fitted values
aov(temperature_C ~ lakename)

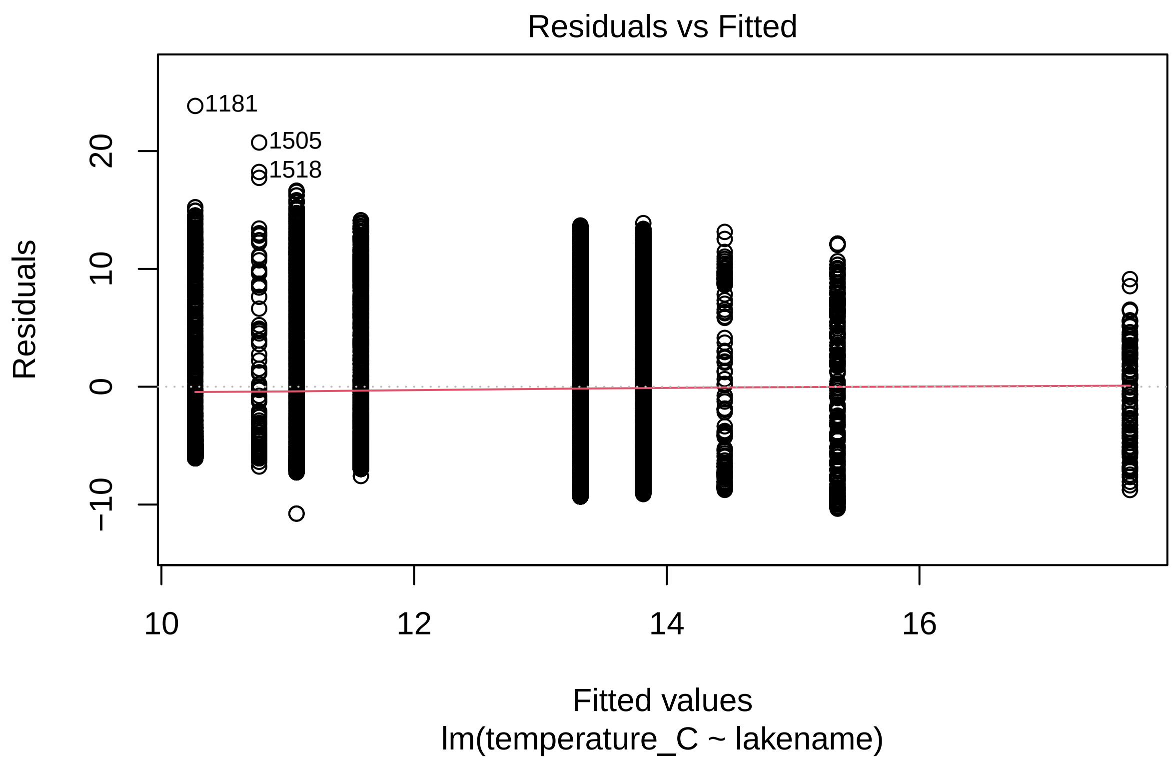


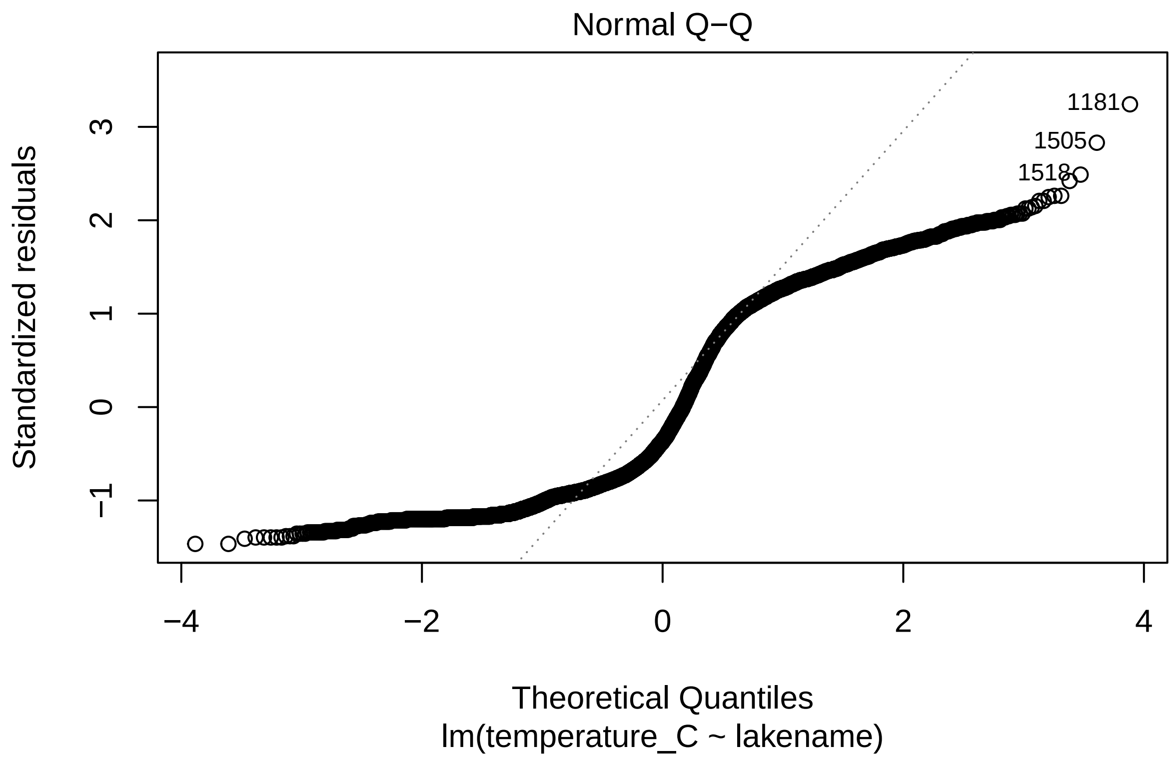

```r
Lake.Temp.Linear <- lm(data = Lake.Temps, temperature_C ~ lakename)
summary(Lake.Temp.Linear)
```

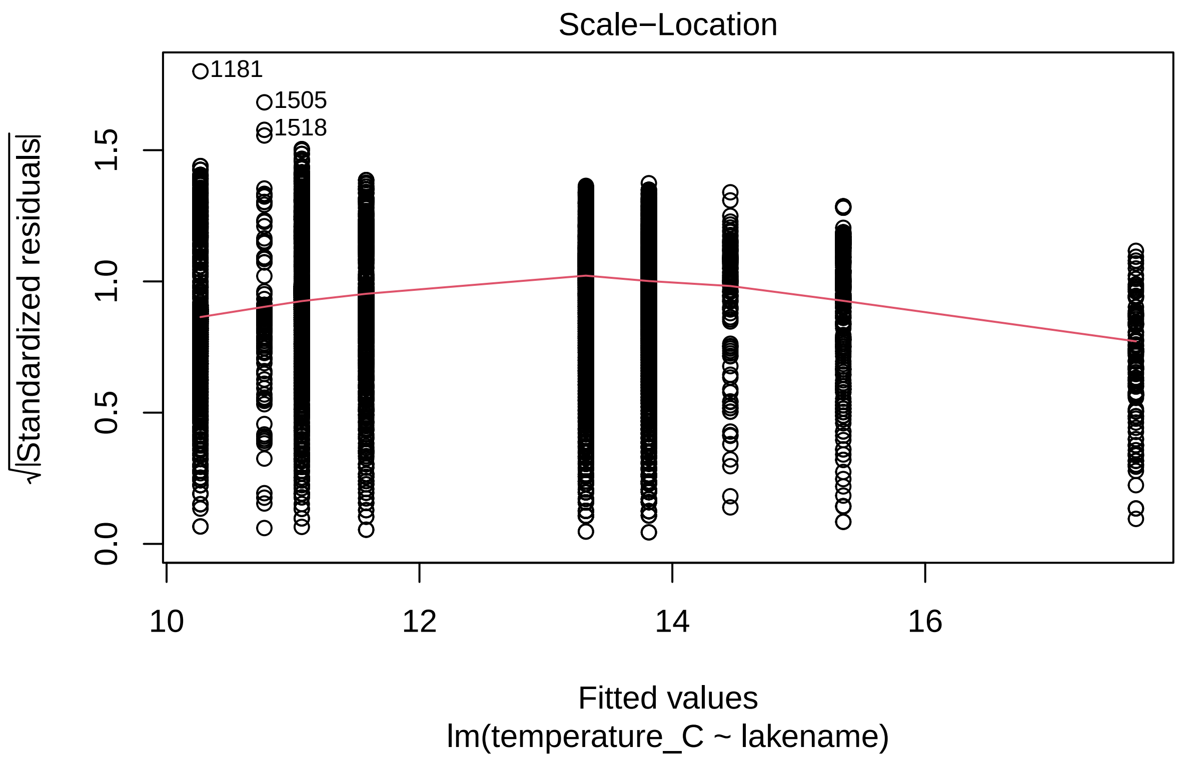
```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = Lake.Temps)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.769  -6.614  -2.679   7.684  23.832
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               17.6664     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
##
```

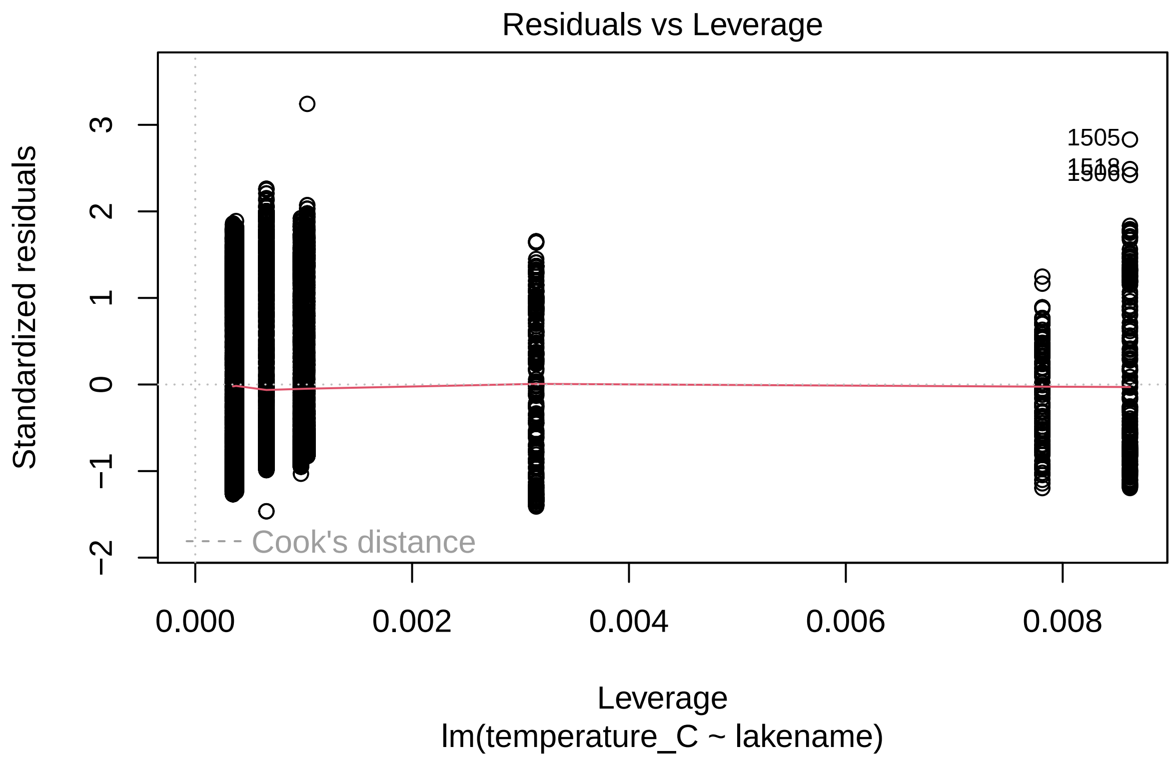
```
## Residual standard error: 7.355 on 9719 degrees of freedom
## Multiple R-squared:  0.03953,    Adjusted R-squared:  0.03874
## F-statistic:    50 on 8 and 9719 DF,  p-value: < 2.2e-16
```

```
plot(Lake.Temp.Linear)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(temperature_C ~ lakename)
1181
1505
1518
−4
−2
0
2
4
−1
0
1
2
3


Scale−Location
1181
1505
1518
√|Standardized residuals|
0.0
0.5
1.0
1.5
10
12
14
16
Fitted values
lm(temperature_C ~ lakename)



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

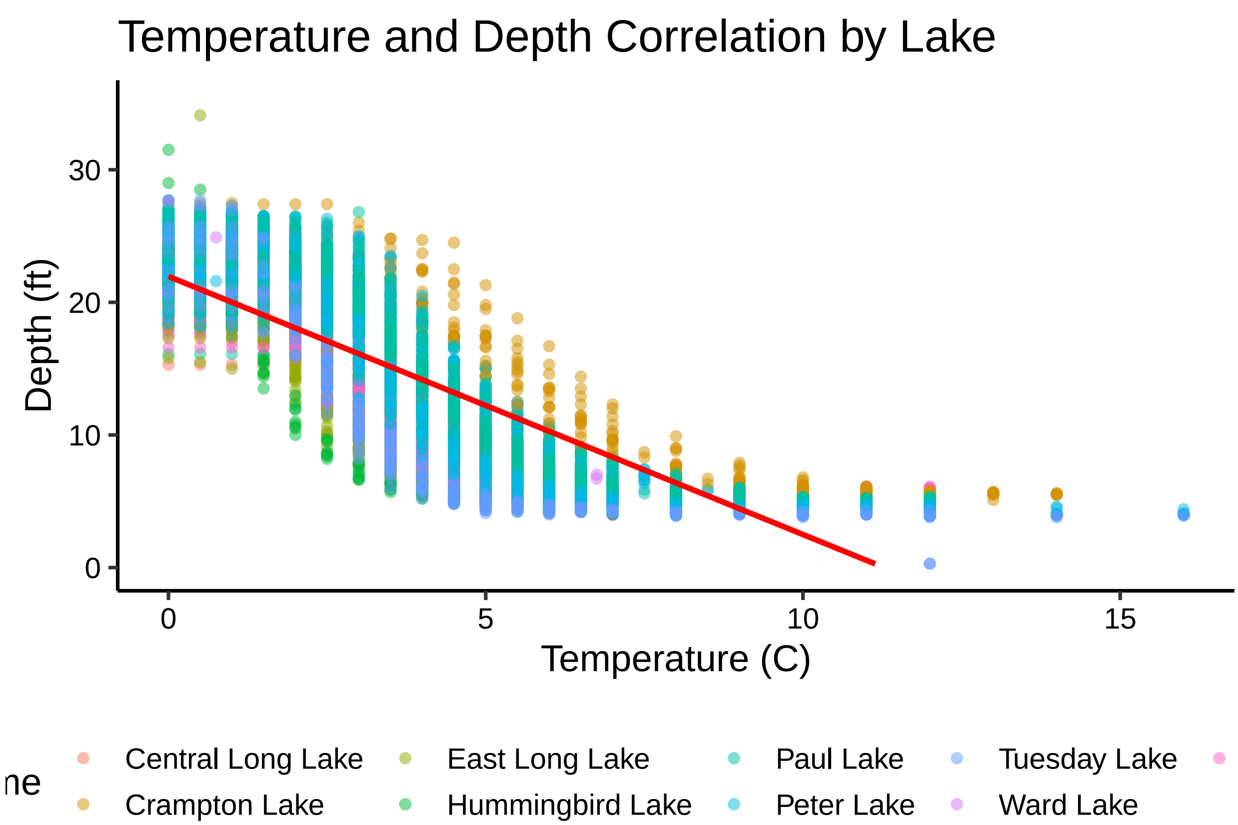
    Answer: For the 8 different lakes evaluated in the anova model, the P value is less than .05 indicating there is a signficatn difference in temperature among all of the lakes.

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. plot temperature change by depth per lake
Temp.Depth.Plot <- ggplot(LakeChem.Pipe, aes(y = temperature_C, x = depth, color = lakename)) +
    geom_point(alpha = 0.5) + ylim(0, 35) + geom_smooth(method = lm, se = FALSE,
    col = "red") + ggtitle("Temperature and Depth Correlation by Lake") + xlab("Temperature (C)") +
    ylab("Depth (ft)")
print(Temp.Depth.Plot)
```

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

```
## Warning: Removed 24 rows containing missing values (geom_smooth).
```



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

```
# 15
TukeyHSD(Lake.Temp_Anova)
```

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

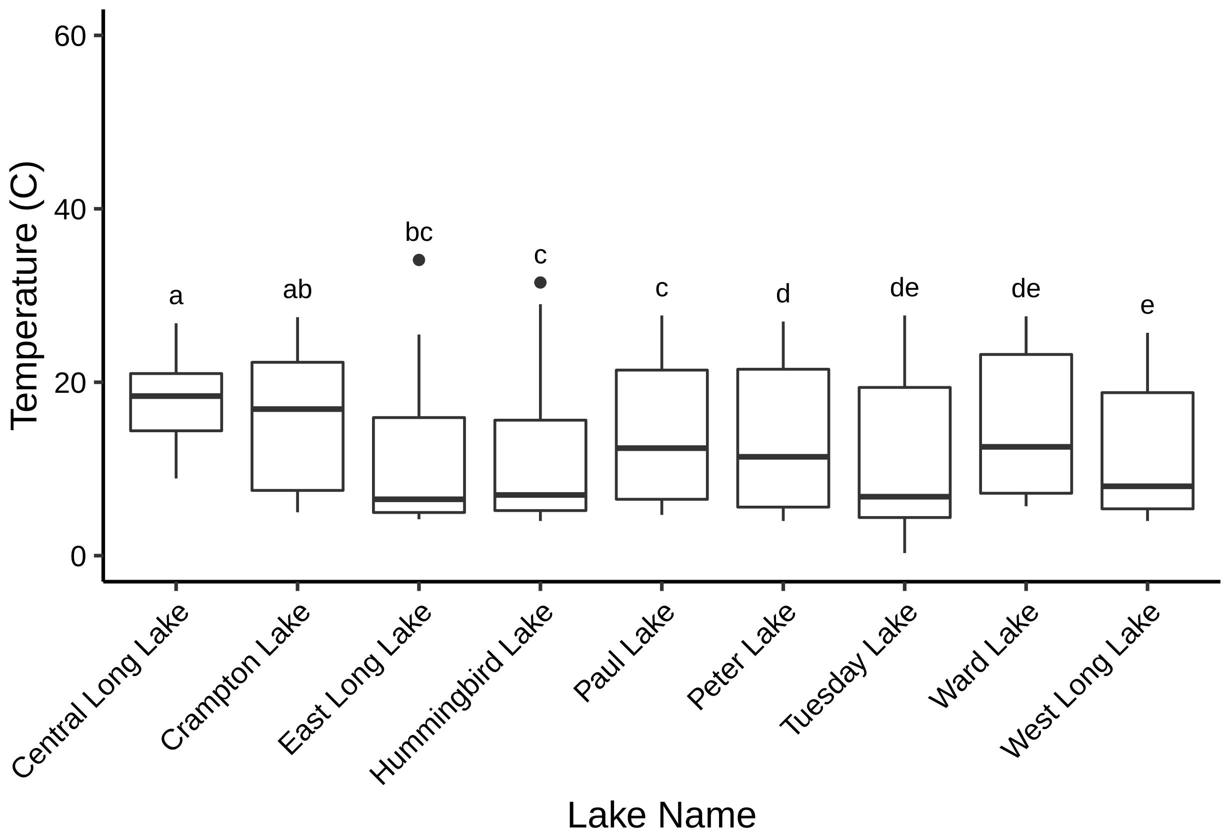
```
## Ward Lake-Crampton Lake           -0.8932661 -3.3684639  1.5819317 0.9714459
## West Long Lake-Crampton Lake      -3.7732318 -5.2378351 -2.3086285 0.0000000
## Hummingbird Lake-East Long Lake    0.5056106 -1.7364925  2.7477137 0.9988050
## Paul Lake-East Long Lake           3.5465903  2.6900206  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
```

```r
# group lakes
Temp.groups <- HSD.test(Lake.Temp_Anova, "lakename", group = TRUE)
Temp.groups
```

```
## $statistics
##   MSerror   Df     Mean       CV
##   54.1016 9719 12.72087 57.82135
##
## $parameters
##    test   name.t ntr StudentizedRange alpha
##   Tukey lakename   9         4.387504  0.05
##
## $means
##                    temperature_C      std    r Min  Max    Q25   Q50    Q75
## Central Long Lake       17.66641 4.196292  128 8.9 26.8 14.400 18.40 21.000
## Crampton Lake           15.35189 7.244773  318 5.0 27.5  7.525 16.90 22.300
## East Long Lake          10.26767 6.766804  968 4.2 34.1  4.975  6.50 15.925
## Hummingbird Lake        10.77328 7.017845  116 4.0 31.5  5.200  7.00 15.625
## Paul Lake               13.81426 7.296928 2660 4.7 27.7  6.500 12.40 21.400
## Peter Lake              13.31626 7.669758 2872 4.0 27.0  5.600 11.40 21.500
## Tuesday Lake            11.06923 7.698687 1524 0.3 27.7  4.400  6.80 19.400
## Ward Lake               14.45862 7.409079  116 5.7 27.6  7.200 12.55 23.200
## West Long Lake          11.57865 6.980789 1026 4.0 25.7  5.400  8.00 18.800
##
## $comparison
## NULL
##
## $groups
##                    temperature_C groups
## Central Long Lake       17.66641      a
```

```
## Crampton Lake          15.35189     ab
## Ward Lake              14.45862     bc
## Paul Lake              13.81426      c
## Peter Lake             13.31626      c
## West Long Lake         11.57865      d
## Tuesday Lake           11.06923     de
## Hummingbird Lake       10.77328     de
## East Long Lake         10.26767      e
##
## attr(,"class")
## [1] "group"
```

```
# graph grouped lakes
Temp.Groups.Plot <- ggplot(LakeChem.Pipe, aes(x = lakename, y = temperature_C)) +
    geom_boxplot() + theme(axis.text.x = element_text(angle = 45, hjust = 1)) + stat_summary(geom = "te
    fun = max, vjust = -1, size = 3.5, label = c("a", "ab", "bc", "c", "c", "d",
        "de", "de", "e")) + labs(x = "Lake Name", y = "Temperature (C)") + ylim(0,
    60)
print(Temp.Groups.Plot)
```



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 grouping, Peter Lake has the same mean temperatures as all the lakes assigned the letter c. This means that Ward and Paul Lakes both have the same mean temperature as

Peter Lake. Statistically speaking, no lakes had temperatures that were different from all other lakes. Each lake was assigned to a group that was statitcially similar to at least one other lake.

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: To determine if Peter and Paul Lake have statistically different mean temperatures, we could perform a T-test. T-tests can be used to compare a variable for two independent samples, in this case, two different lakes.

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?

```
# filter for Crampton and Ward Lakes
Lake.Filter <- filter(LakeChem.Pipe, lakename == "Crampton Lake" | lakename == "Ward Lake")

# run t-tests
twosample <- t.test(Lake.Filter$temperature_C ~ Lake.Filter$lakename)
twosample
```

```
##
##  Welch Two Sample t-test
##
## data:  Lake.Filter$temperature_C by Lake.Filter$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 r
## 95 percent confidence interval:
##  -0.6821129  2.4686451
## sample estimates:
## mean in group Crampton Lake     mean in group Ward Lake
##                    15.35189                    14.45862
```

```
twosample2 <- lm(Lake.Filter$temperature_C ~ Lake.Filter$lakename)
summary(twosample2)
```

```
##
## Call:
## lm(formula = Lake.Filter$temperature_C ~ Lake.Filter$lakename)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3519  -7.5286   0.1947   7.0481  13.1414
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     15.3519     0.4087   37.56   <2e-16 ***
## Lake.Filter$lakenameWard Lake  -0.8933     0.7906   -1.13    0.259
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 7.289 on 432 degrees of freedom
## Multiple R-squared:  0.002946,   Adjusted R-squared:  0.0006383
## F-statistic: 1.277 on 1 and 432 DF,  p-value: 0.2592
```

> Answer: The T-test indicates that there is not a statistical difference between Ward and Crampton Lake. The p-value for the test is .259 which is above .05. This P-value lines up with what we know from the tukey test that grouped both lakes together as having mean temperature values with statistcally similar means.