Assignment 7: GLMs (Linear Regressios, ANOVA, & t-tests)

Claire Pajka

Fall 2023

OVERVIEW

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

Directions

1. Rename this file `<FirstLast>_A07_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
#install.packages("agricolae")
library(tidyverse);library(lubridate);library(here); library(agricolae); library(dplyr)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
## here() starts at C:/Users/cepaj/OneDrive/Documents/EDE_Fall2023
```

```
here()
```

```
## [1] "C:/Users/cepaj/OneDrive/Documents/EDE_Fall2023"
```

```
chemphysics_raw <- read.csv(
  here("Data", "Raw", "NTL-LTER_Lake_ChemistryPhysics_Raw.csv"),
  stringsAsFactors = TRUE)
View(chemphysics_raw)
chemphysics_raw$sampledate <-
  mdy(chemphysics_raw$sampledate)

#2
mytheme <- theme_classic(base_size=12)+
  theme(
    plot.title= element_text(size= 12, color = "darkblue",
                             face="bold", hjust = 0.5),
    axis.text = element_text(size = 12, color = "black"),
    legend.position = 'right',
    legend.background = element_blank(),
    legend.box.background = element_rect(colour = "black"),
    plot.background = element_rect(color = "black"),
    axis.line = element_line(size = 0.65, color = "black"))
```

```
## Warning: The ‘size‘ argument of ‘element_line()‘ is deprecated as of ggplot2 3.4.0.
## i Please use the ‘linewidth‘ argument instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```
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: There is no difference in lake temperatures at different depths across the study lakes. Ha: Lake temperatures vary by depth across the study 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.

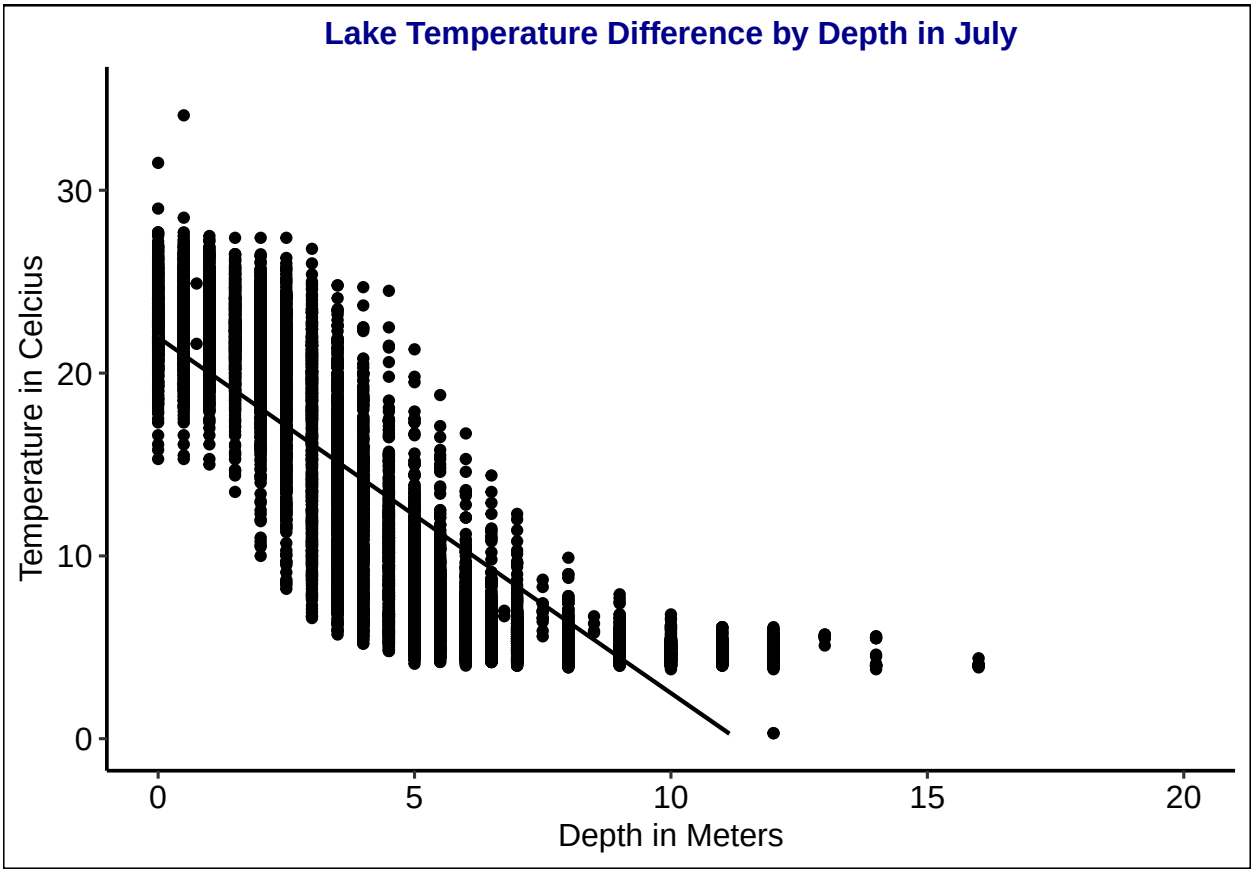
```
#4
subsetdepth_temp <- chemphysics_raw %>%
  #mutate(sampledate = as.Date(sampledate, format = "%m- %d - %Y"),
         #Month = month(sampledate),
         #Year = year(sampledate))%>%
```

```
  filter(month(chemphysics_raw$sampledate)== 7) %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  na.omit()
  view(subsetdepth_temp)

#5
Depth_temp_plot <-
  ggplot(subsetdepth_temp,
         aes(
           y=temperature_C,
           x= depth))+
  geom_point()+
  mytheme+
  geom_smooth(method = lm, linewidth = .75, color = "black")+
  xlim(0,20)+
  ylim(0, 35)+
  labs(title= "Lake Temperature Difference by Depth in July",
       fontface = "bold",
       color = "Lake Name")+
  xlab("Depth in Meters")+
  ylab("Temperature in Celcius")
print(Depth_temp_plot)
```

```
## `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: Generally, the graph shows that as depth increases, the temperature in degrees Celcius decreases. The data appears to decrease in a nonlinear manner (as demonstrated by its difference from the trend line), similar to a logarithmic curve. This suggests that the relationship between temperature in Celcius and depth is nonlinear / the response of temperature by depth is nonlinear.

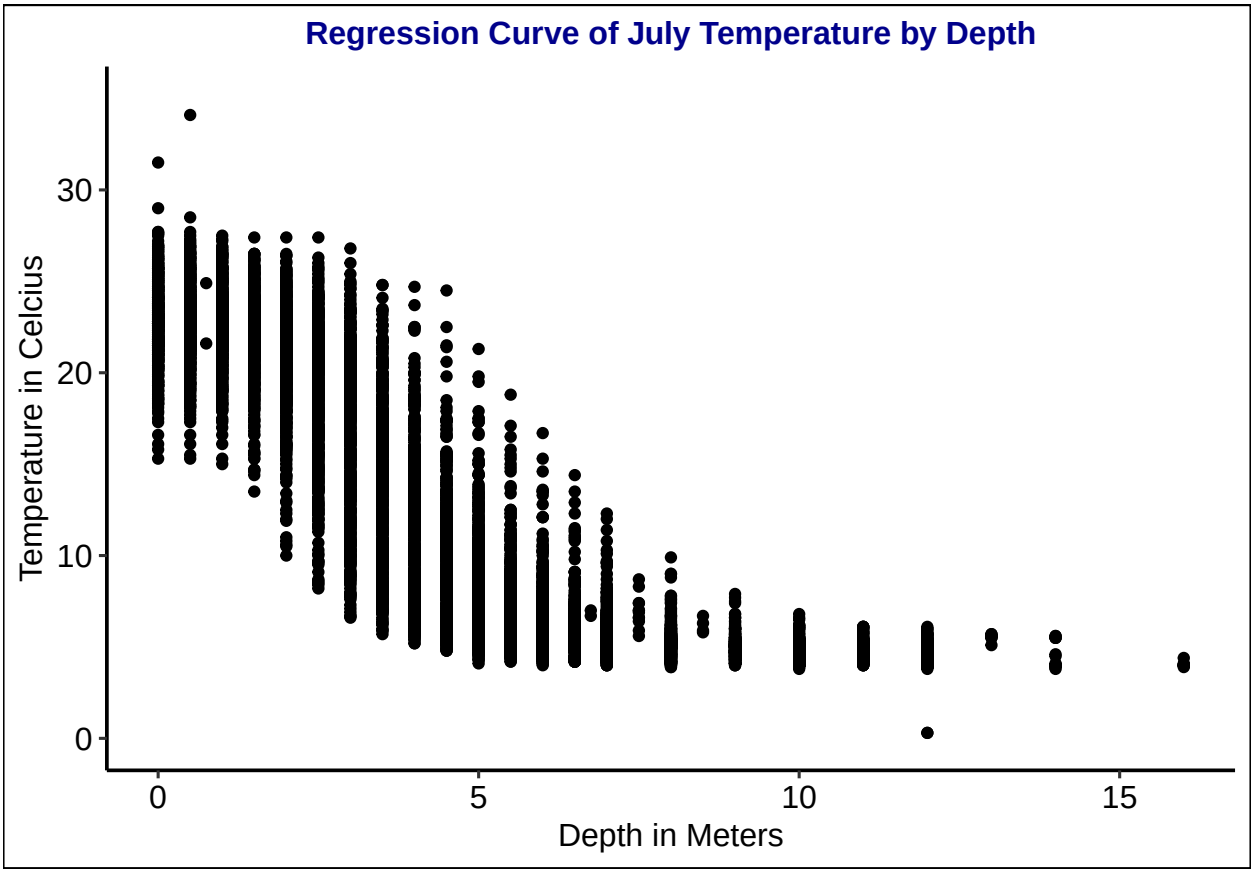
7. Perform a linear regression to test the relationship and display the results

```
#7
depth_temp_regression <-
  lm(subsetdepth_temp$depth ~
       subsetdepth_temp$temperature_C)
summary(depth_temp_regression)
```

```
##
## Call:
## lm(formula = subsetdepth_temp$depth ~ subsetdepth_temp$temperature_C)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0685 -1.1065 -0.2334  0.9668  8.0964
##
```

```
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     9.573728   0.033803   283.2   <2e-16 ***
## subsetdepth_temp$temperature_C -0.379578   0.002289  -165.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.694 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
```

```
depth_temp_regression_plot <-
  ggplot(subsetdepth_temp, aes(x = depth, y = temperature_C))+
  ylim(0,35)+
  geom_point()+
   labs(title= "Regression Curve of July Temperature by Depth",
       fontface = "bold")+
  xlab("Depth in Meters")+
  ylab("Temperature in Celcius")
print(depth_temp_regression_plot)
```



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^2 explains how much variability is explained by depth, so the r^2 value of 0.7387 means that 73.87% of the variability is accounted for by depth. The degrees of freedom are 9726. The p value is less than 2.2e-16, which is less than 0.05, so we would call this result statistically significant.For every 1 meter increase in depth (going lower into the water), the temperature is going to decrease by 0.379578 degrees, and every one meter decrease in depth (going closer to the surface of the water) the temperature will increase by 0.379578 degrees Celcius.

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
TempbyAll.regression <- lm(data = subsetdepth_temp, temperature_C ~ year4 + daynum + depth)
TempbyAll.regression
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = subsetdepth_temp)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##    -8.57556      0.01134      0.03978     -1.94644
```

```
step(TempbyAll.regression)
```

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

```r
print(TempbyAll.regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = subsetdepth_temp)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##    -8.57556      0.01134      0.03978     -1.94644
```

```r
#10
Temp_by_depthmultiple <- lm(data = subsetdepth_temp, temperature_C ~ year4 + daynum + depth)
summary(Temp_by_depthmultiple)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = subsetdepth_temp)
##
## 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 variables includes year4, daynum, and depth. For the multiple regression with the variables suggested by the AIC method, the multiple r squared is 0.7412 (adjusted 0.7411), meaning that 74.12% of the variance is accounted for by the model. The single linear regression of only depth and temperature provided an r squared of 0.7387, which means that the adding in more explanatory variables very slightly increased the amount of the observed variance that the model explains.

Analysis of Variance

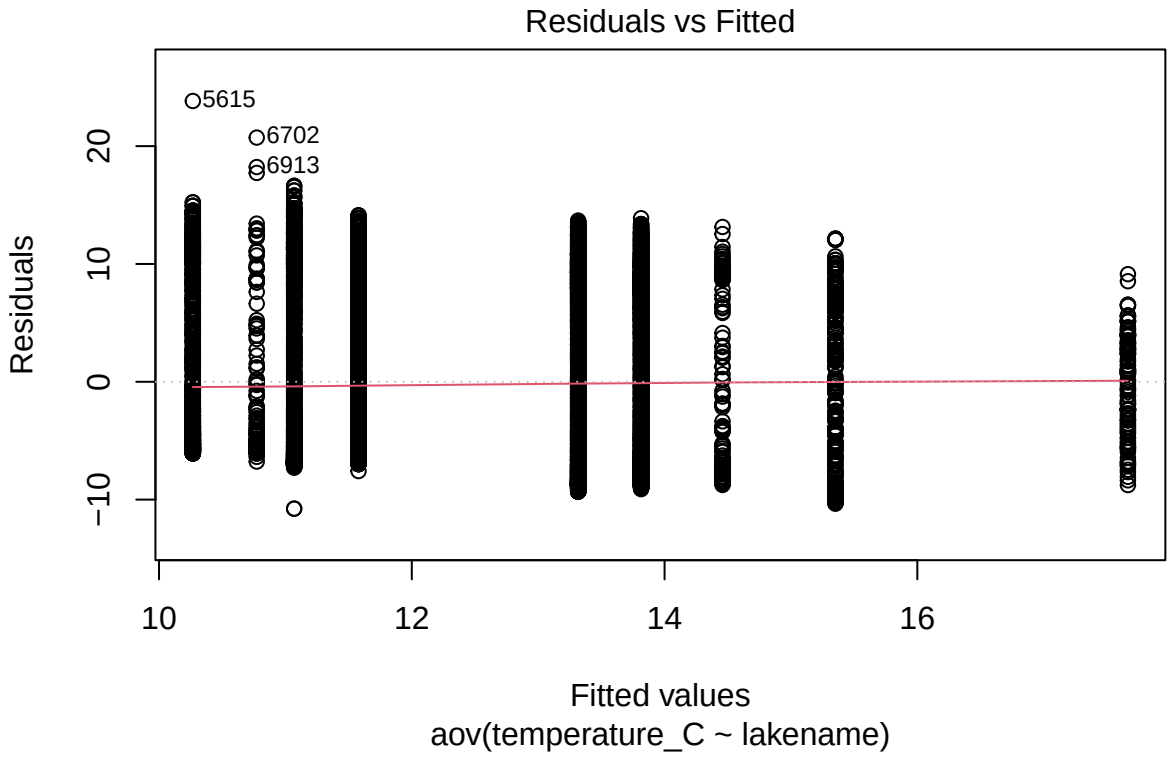
12. Now we want to see whether the different lakes have, on average, different temperatures in the month of July. Run an ANOVA test to complete this analysis. (No need to test assumptions of normality

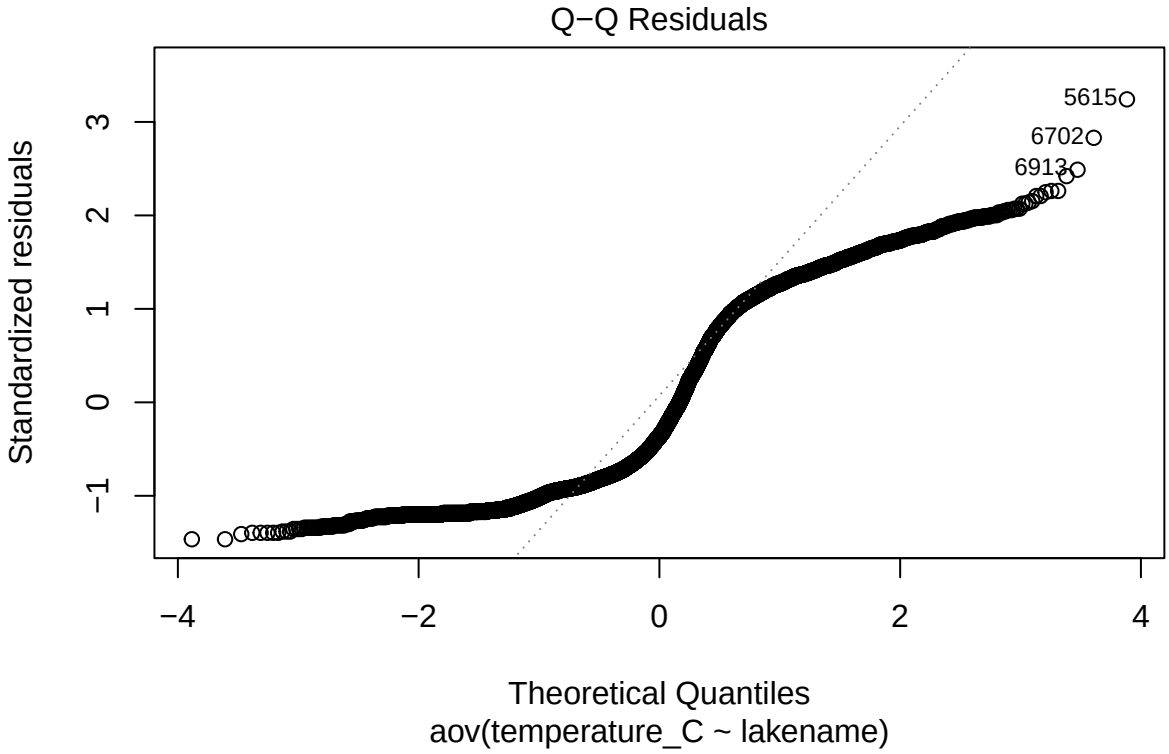
or similar variances.) Create two sets of models: one expressed as an ANOVA models and another expressed as a linear model (as done in our lessons).

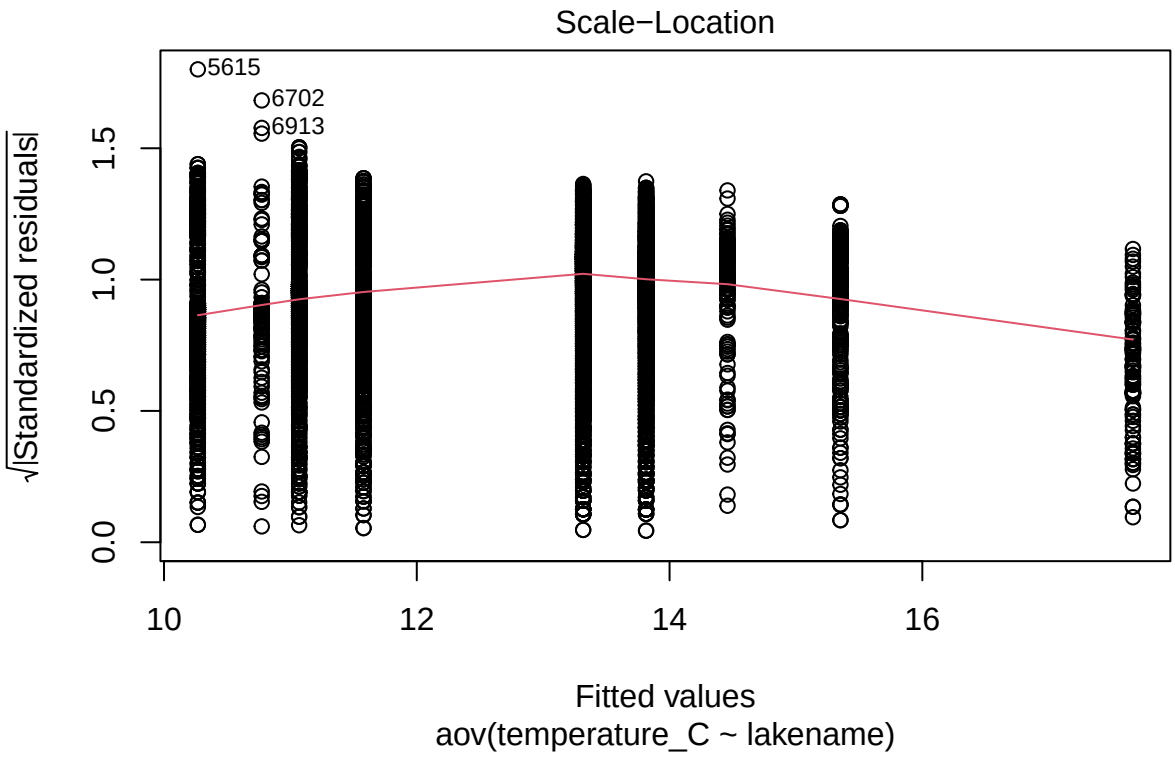
```
#12
July_temp_anova <- aov(data = subsetdepth_temp, temperature_C ~ lakename)
summary(July_temp_anova)
```

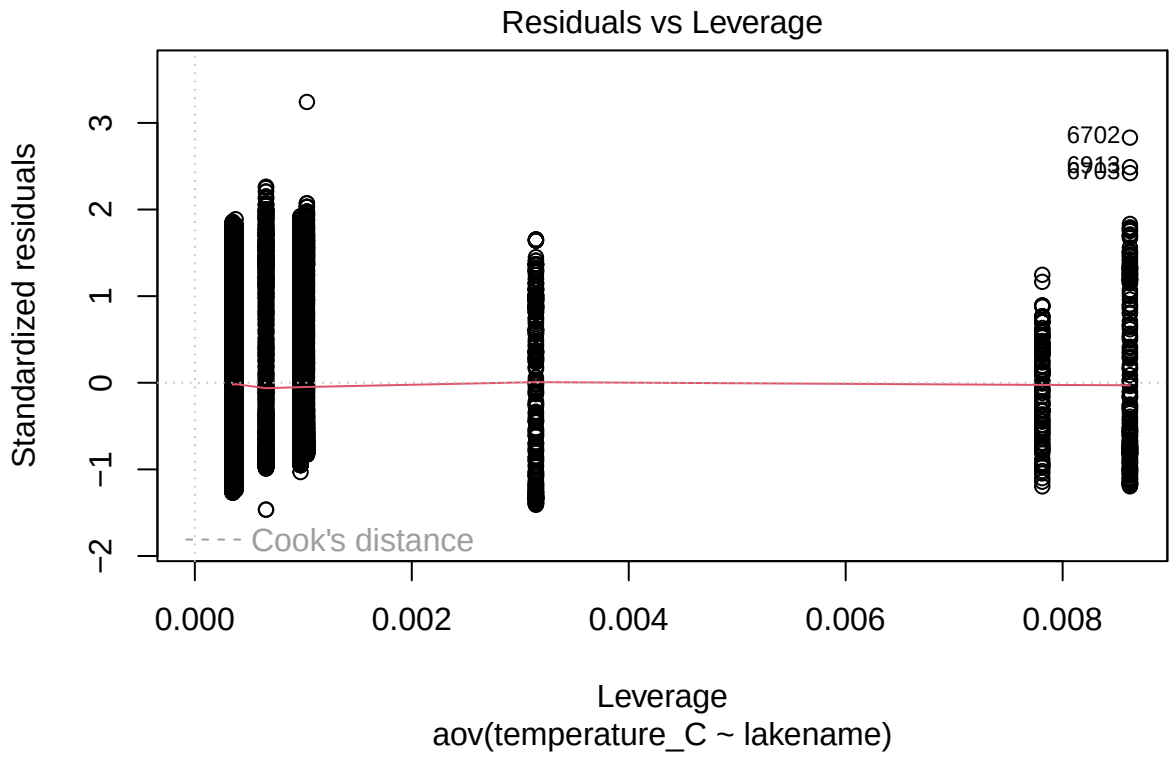
```
##               Df Sum Sq Mean Sq F value Pr(>F)
## lakename       8  21642  2705.2      50 <2e-16 ***
## Residuals   9719 525813    54.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(July_temp_anova)
```




Q−Q Residuals
Standardized residuals
−1
0
1
2
3
5615
6702
6913
−4
−2
0
2
4
Theoretical Quantiles
aov(temperature_C ~ lakename)


Scale−Location
5615
6702
6913
√|Standardized residuals|
0.0
0.5
1.0
1.5
10
12
14
16
Fitted values
aov(temperature_C ~ lakename)

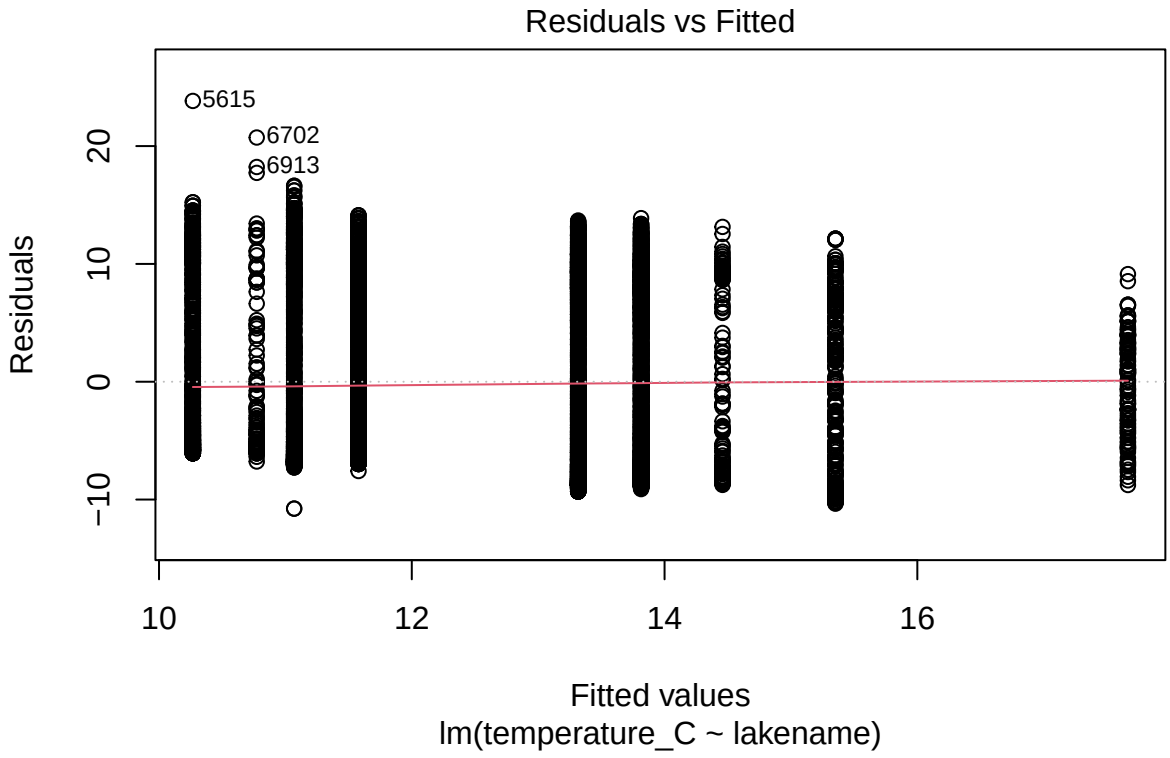


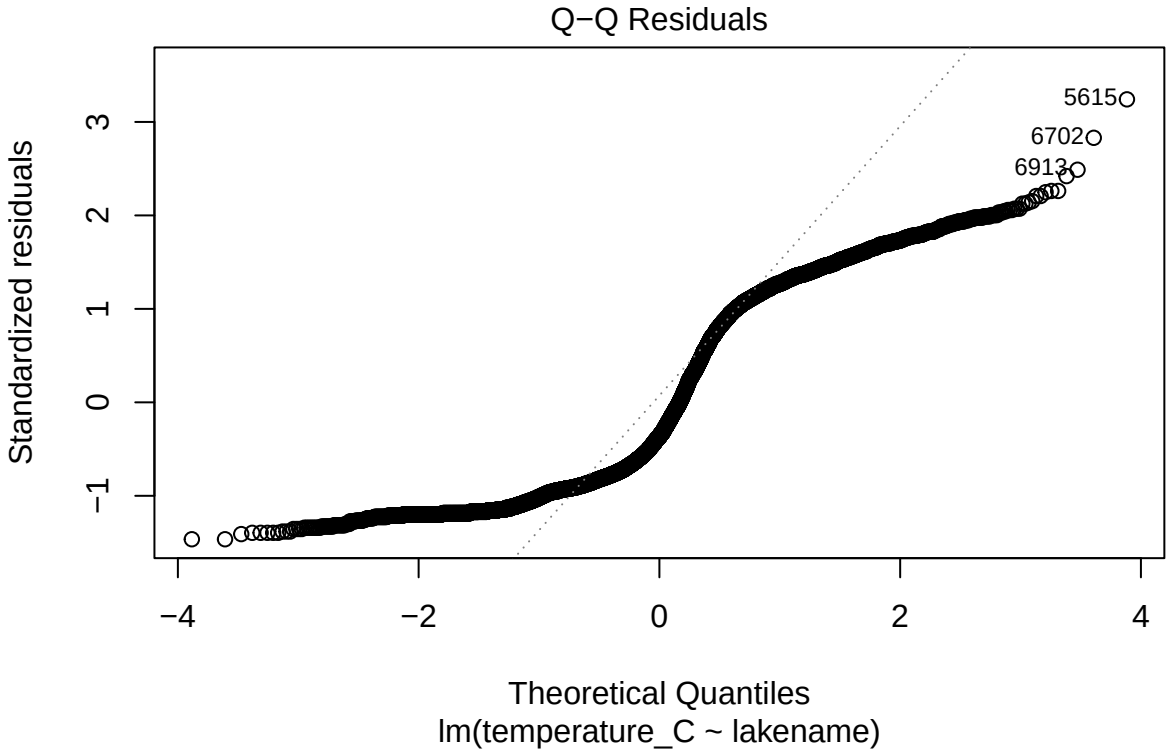

```
July_temp_anova2 <- lm(data = subsetdepth_temp, temperature_C ~ lakename)
summary(July_temp_anova2)
```

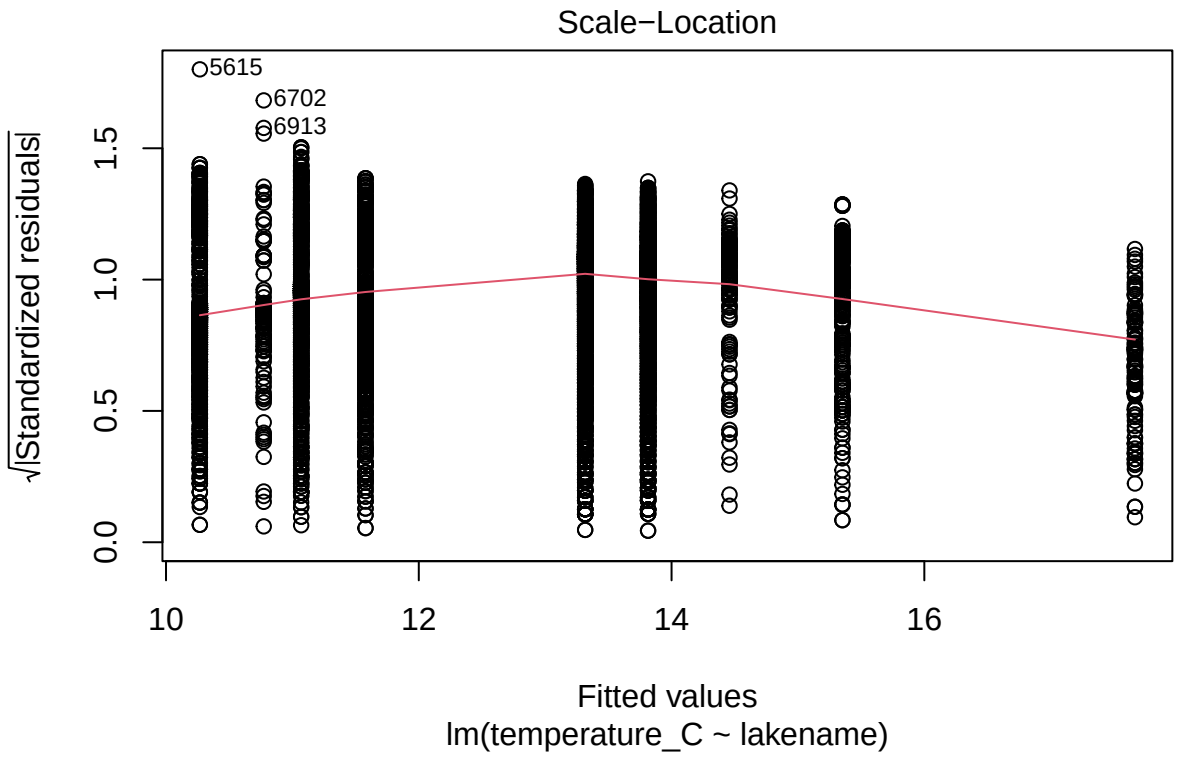
```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = subsetdepth_temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.769  -6.614  -2.679   7.684  23.832
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)             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

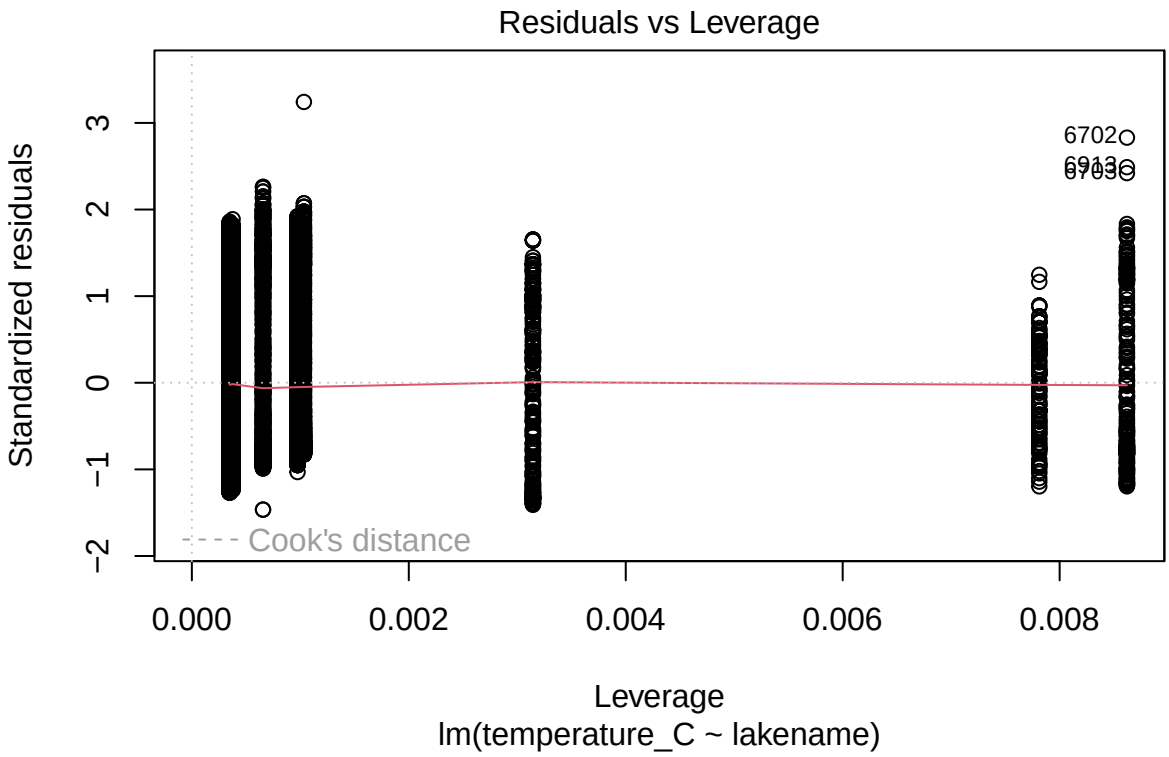
```
## Residual standard error: 7.355 on 9719 degrees of freedom
## Multiple R-squared:  0.03953,    Adjusted R-squared:  0.03874
## F-statistic:    50 on 8 and 9719 DF,  p-value: < 2.2e-16
```

```
plot(July_temp_anova2)
```




Q−Q Residuals
Standardized residuals
−1
0
1
2
3
5615
6702
6913
−4
−2
0
2
4
Theoretical Quantiles
lm(temperature_C ~ lakename)


Scale−Location
5615
6702
6913
√|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: Given that the p value is less than 0.05, we reject the null hypothesis that there is no significant difference between mean temperature among the lakes. The data implies that there is a significant difference in mean temperature among the lakes.

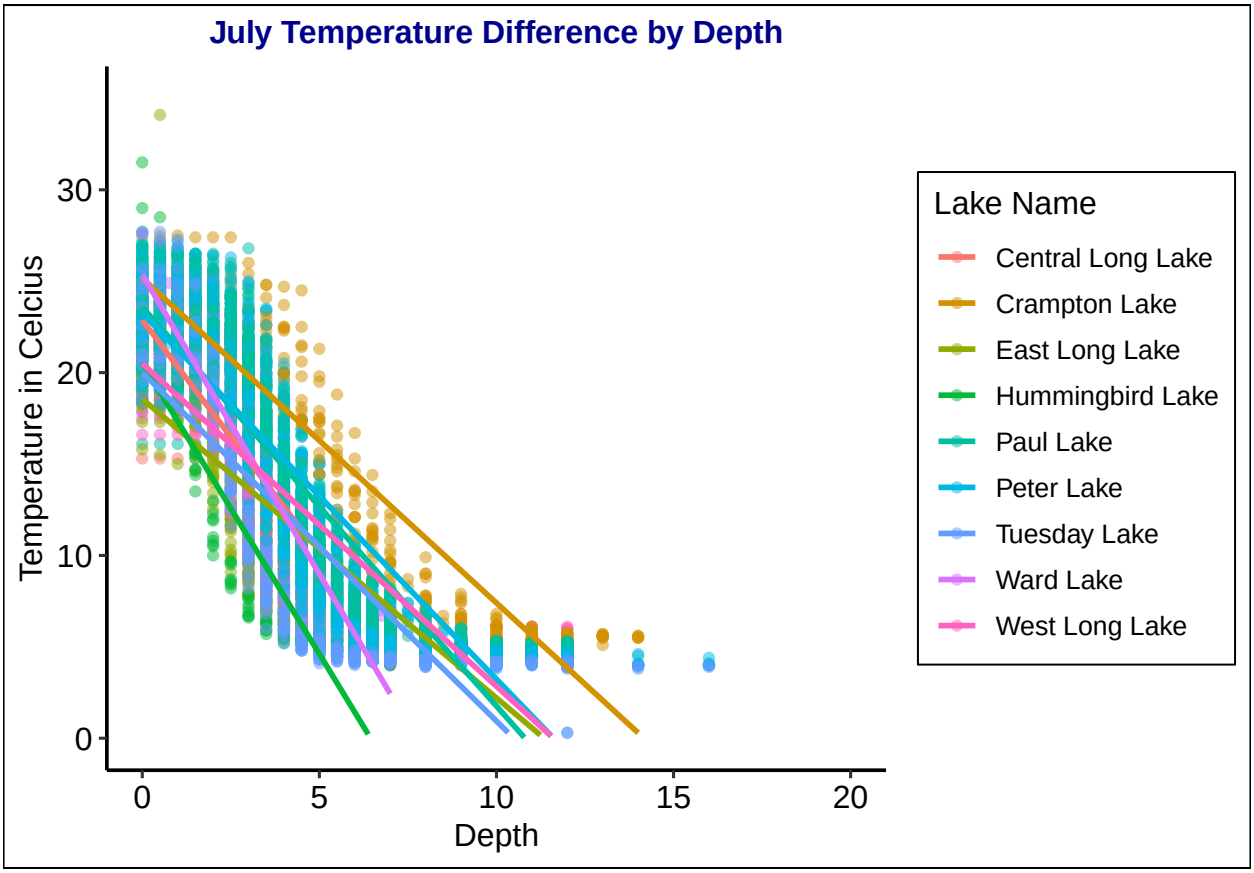
14. Create a graph that depicts temperature by depth, with a separate color for each lake. Add a geom_smooth (method = "lm", se = FALSE) for each lake. Make your points 50 % transparent. Adjust your y axis limits to go from 0 to 35 degrees. Clean up your graph to make it pretty.

```
#14.
Depth_temp_plot2 <-
  ggplot(subsetdepth_temp,
         aes(
           y=temperature_C,
           x= depth,
           color = lakename))+
  geom_point(alpha= 0.5)+
  mytheme+
  geom_smooth(method = lm, se= FALSE)+
  xlim(0,20)+
  ylim(0, 35)+
  labs(title= "July Temperature Difference by Depth",
       fontface = "bold",
       color = "Lake Name")+
```

```
  xlab("Depth")+
  ylab("Temperature in Celcius")
print(Depth_temp_plot2)
```

```
## `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
Temp.groups <- HSD.test(July_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        se Min  Max    Q25   Q50
## Central Long Lake      17.66641 4.196292  128 0.6501298 8.9 26.8 14.400 18.40
## Crampton Lake          15.35189 7.244773  318 0.4124692 5.0 27.5  7.525 16.90
## East Long Lake         10.26767 6.766804  968 0.2364108 4.2 34.1  4.975  6.50
## Hummingbird Lake       10.77328 7.017845  116 0.6829298 4.0 31.5  5.200  7.00
## Paul Lake              13.81426 7.296928 2660 0.1426147 4.7 27.7  6.500 12.40
## Peter Lake             13.31626 7.669758 2872 0.1372501 4.0 27.0  5.600 11.40
## Tuesday Lake           11.06923 7.698687 1524 0.1884137 0.3 27.7  4.400  6.80
## Ward Lake              14.45862 7.409079  116 0.6829298 5.7 27.6  7.200 12.55
## West Long Lake         11.57865 6.980789 1026 0.2296314 4.0 25.7  5.400  8.00
##                      Q75
## Central Long Lake 21.000
## Crampton Lake     22.300
## East Long Lake    15.925
## Hummingbird Lake  15.625
## Paul Lake         21.400
## Peter Lake        21.500
## Tuesday Lake      19.400
## Ward Lake         23.200
## West Long Lake    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"
```

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: Statistically speaking, Paul Lake has the same mean as Peter Lake. There are no lakes that have a mean temperature that is statistically distinct from all of the other lakes: all of the lakes have at least one other lake with which the means are the same.

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: We could use a two sample t-test in order to see if Peter and Paul Lake have different means.Two sample t-tests are used to compare the means of two independent populations and test if the difference between them is statistically significant.

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 <- subsetdepth_temp %>%
  filter(lakename == "Crampton Lake"| lakename =="Ward Lake")
view(crampton_ward)

july_temp_twosample <- t.test(crampton_ward$temperature_C ~ crampton_ward$lakename)
july_temp_twosample
```

```
##
##  Welch Two Sample t-test
##
## data:  crampton_ward$temperature_C by crampton_ward$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: Running the two sample t test, we get a p value of 0.2649. Because this is greater than 0.05, we fail to reject the null. The null hypothesis of the two sample t test is that the means are the same, and because we failed to reject the null, the data asserts that the mean July temperatures for Cramptom Lake and Ward Lake are the same. This matches with the answer in #16, because we saw that Crampton Lake had a grouping of ab, and Ward Lake had a grouping of bc. Because they have the group 'b' in common, question 16 showed that there would not be a statistical difference between the means of Crampton and Ward Lake.