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

Student Name

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
library(tidyverse)
```

```
## -- 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 1.9.3     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
```

```
library(agricolae)
library(here)
```

```
## here() starts at /Users/vincient/Desktop/Projects/EDA_class
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(dplyr)
here()
```

```
## [1] "/Users/vincient/Desktop/Projects/EDA_class"
```

```
ChemPhysics <- read.csv(here("Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv"), stringsAsFactors = TRU

ChemPhysics$sampledate <- as.Date(ChemPhysics$sampledate, format = "%m/%d/%Y")

#2

mytheme <- theme_classic(base_size = 15) +
  theme(axis.text = element_text(color = "blue"),
        legend.position = "top")
theme_set(mytheme)
```

Simple regression

Our first research question is: Does mean lake temperature recorded during July change with depth across all lakes?

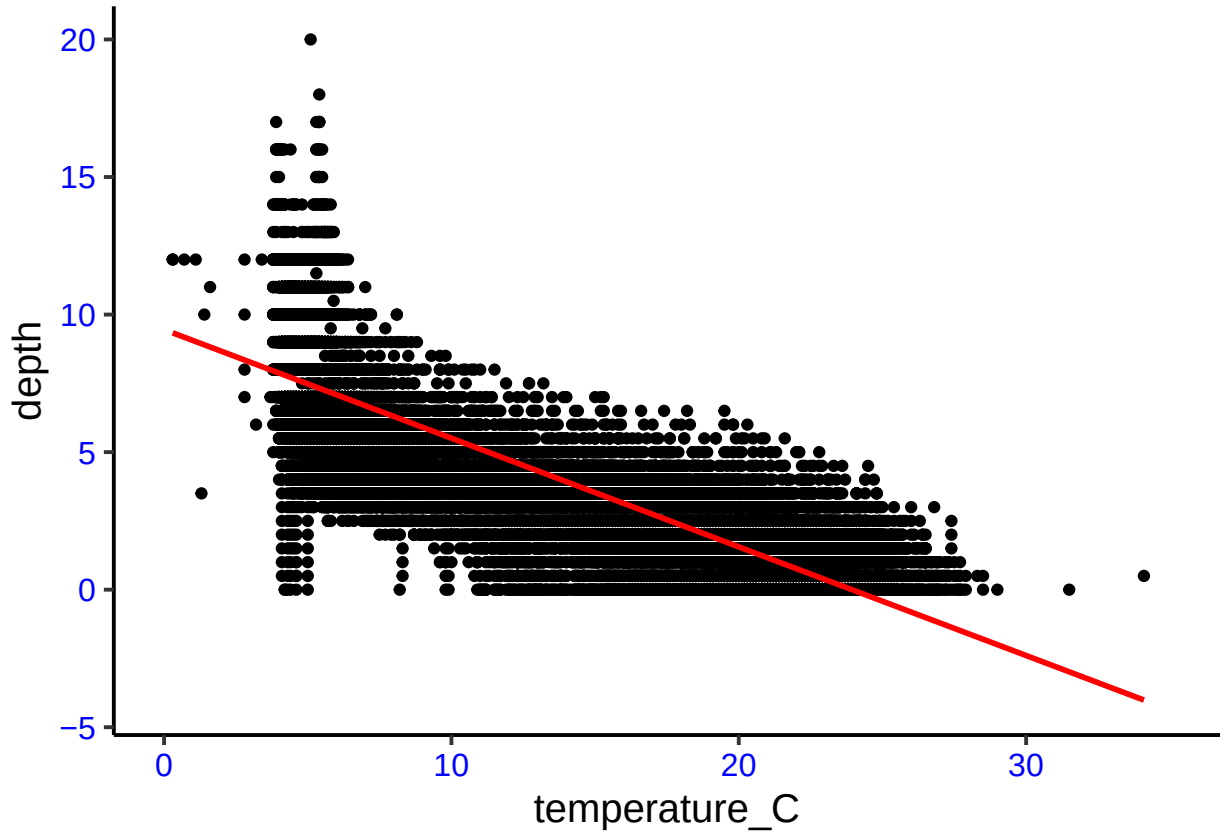
3. State the null and alternative hypotheses for this question: > Answer: H0: mean lake temperature recorded during July change with depth across all lakes. Ha: mean lake temperature recorded during July do not change with depth across all lakes.
4. Wrangle your NTL-LTER dataset with a pipe function 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
ChemPhysics.july <- ChemPhysics %>%
  filter(daynum == "07") %>%
  select(lakename,year4,daynum,depth,temperature_C) %>%
  drop_na()

#5
TempByDepth.plot <-
  ggplot(ChemPhysics, aes( x=temperature_C,
                           y=depth))+
                          xlim(0,35)+
  geom_point()+
  geom_smooth(method = "lm",col= "red")
print(TempByDepth.plot)
```

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

## Warning: Removed 3858 rows containing non-finite values (`stat_smooth()`).

## Warning: Removed 3858 rows containing missing values (`geom_point()`).
```



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 box plot indicates that as depth decreases our temperature will increase. The distribution of points lets us know that significance evidence of this.

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

```
#7
ChemPhysics.regression <-
  lm(ChemPhysics$temperature_C ~
       ChemPhysics$depth)
summary(ChemPhysics.regression)
```

```
##
## Call:
## lm(formula = ChemPhysics$temperature_C ~ ChemPhysics$depth)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7864  -3.1363  -0.1219   3.1815  19.2568
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       19.986395   0.037166   537.8   <2e-16 ***
## ChemPhysics$depth -1.707162   0.006366  -268.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.961 on 34754 degrees of freedom
##   (3858 observations deleted due to missingness)
## Multiple R-squared:  0.6742, Adjusted R-squared:  0.6742
## F-statistic: 7.192e+04 on 1 and 34754 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:This model demostrates a negative correlation, as depth decreasing the temparature is increasing.The p-value is also less than .05 which means that we should look further intot this. With 34754 degrees of freedom and a r-squared value that tells us that depth is explains about 67% of our variablity in temparature. The P-Value is lower than our confidnce level, concluding that this is a meaningful regression.

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

ChemPhysicsNew.regression <- lm(
  data = ChemPhysics,
  temperature_C ~ depth + year4+ daynum )
summary(ChemPhysicsNew.regression)
```

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

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.7228  -2.8606  -0.1706   2.9267  17.8338
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -2.2680808  4.5365880   -0.500 0.617111
## depth       -1.7086514  0.0060824 -280.915  < 2e-16 ***
## year4        0.0077342  0.0022622    3.419 0.000629 ***
## daynum       0.0349904  0.0006083   57.517  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.784 on 34752 degrees of freedom
##   (3858 observations deleted due to missingness)
## Multiple R-squared:  0.7026, Adjusted R-squared:  0.7026
## F-statistic: 2.737e+04 on 3 and 34752 DF,  p-value: < 2.2e-16
```

```
step(ChemPhysicsNew.regression)
```

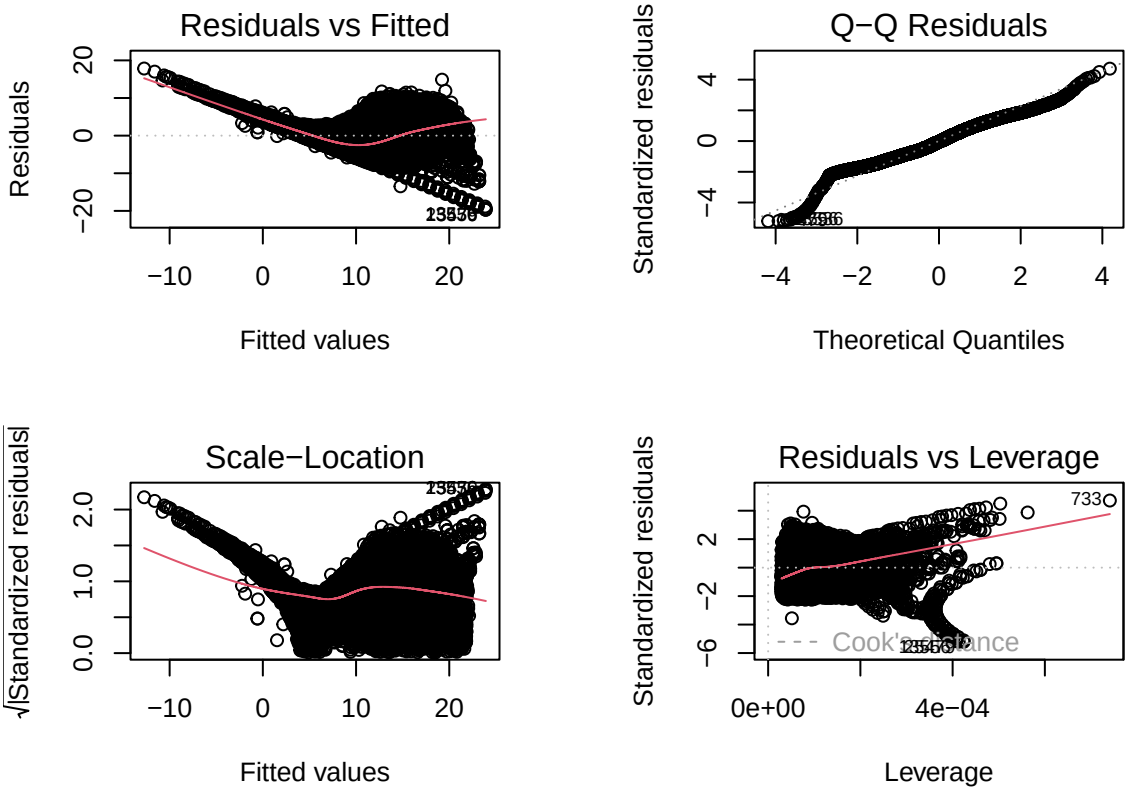
```
## Start:  AIC=92515.66
## temperature_C ~ depth + year4 + daynum
##
##          Df Sum of Sq     RSS    AIC
## <none>                 497693  92516
## - year4   1       167  497861  92525
## - daynum  1     47378  545071  95674
## - depth   1   1130140 1627834 133700

##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = ChemPhysics)
##
## Coefficients:
## (Intercept)        depth        year4       daynum
##   -2.268081    -1.708651     0.007734     0.034990
```

```
CPmodel <- lm(data = ChemPhysics, temperature_C ~ depth + year4 + daynum)
summary(CPmodel)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = ChemPhysics)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.7228  -2.8606  -0.1706   2.9267  17.8338
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -2.2680808  4.5365880   -0.500 0.617111
## depth       -1.7086514  0.0060824 -280.915  < 2e-16 ***
```

```
## year4        0.0077342  0.0022622    3.419 0.000629 ***
## daynum       0.0349904  0.0006083   57.517  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.784 on 34752 degrees of freedom
##   (3858 observations deleted due to missingness)
## Multiple R-squared:  0.7026, Adjusted R-squared:  0.7026
## F-statistic: 2.737e+04 on 3 and 34752 DF,  p-value: < 2.2e-16
```

```
#10

par(mfrow = c(2,2), mar=c(4,4,4,4))
plot(ChemPhysicsNew.regression)
```




```
par(mfrow = c(1,1))
```

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 sugest that we remove none of the varibles,our AIC is lowestest with all of them present.But also in this we an observed variance of 70% which is a 3 percent increase from our earlier variance.

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
ChemPhysics.july <- ChemPhysics %>%
  group_by(lakeid, lakename, sampledate)%>%
  summarise(temperature_C = sum(temperature_C))
```

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

```
CP.Totals.anova <- aov(data = ChemPhysics.july, temperature_C ~ lakename)
summary(CP.Totals.anova)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## lakename     6  55205    9201   3.075 0.0102 *
## Residuals   67 200504    2993
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 1828 observations deleted due to missingness
```

```
CP.Totals.anova2 <- lm(data = ChemPhysics.july, temperature_C ~ lakename)
summary(CP.Totals.anova2)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = ChemPhysics.july)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -104.883  -39.121    4.256   41.939  101.950
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  211.15      38.68   5.459 7.54e-07 ***
## lakenameEast Long Lake       -96.92      49.94  -1.941   0.0565 .
## lakenameHummingbird Lake     -62.82      49.94  -1.258   0.2128
## lakenamePaul Lake            -50.15      41.03  -1.222   0.2259
## lakenamePeter Lake           -16.27      39.95  -0.407   0.6852
## lakenameTuesday Lake         -73.06      40.77  -1.792   0.0777 .
## lakenameWest Long Lake       -82.70      54.70  -1.512   0.1353
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.7 on 67 degrees of freedom
##   (1828 observations deleted due to missingness)
## Multiple R-squared:  0.2159, Adjusted R-squared:  0.1457
## F-statistic: 3.075 on 6 and 67 DF,  p-value: 0.01018
```

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

    Answer: There is, the max of the anova2 is almost half the average anova1.

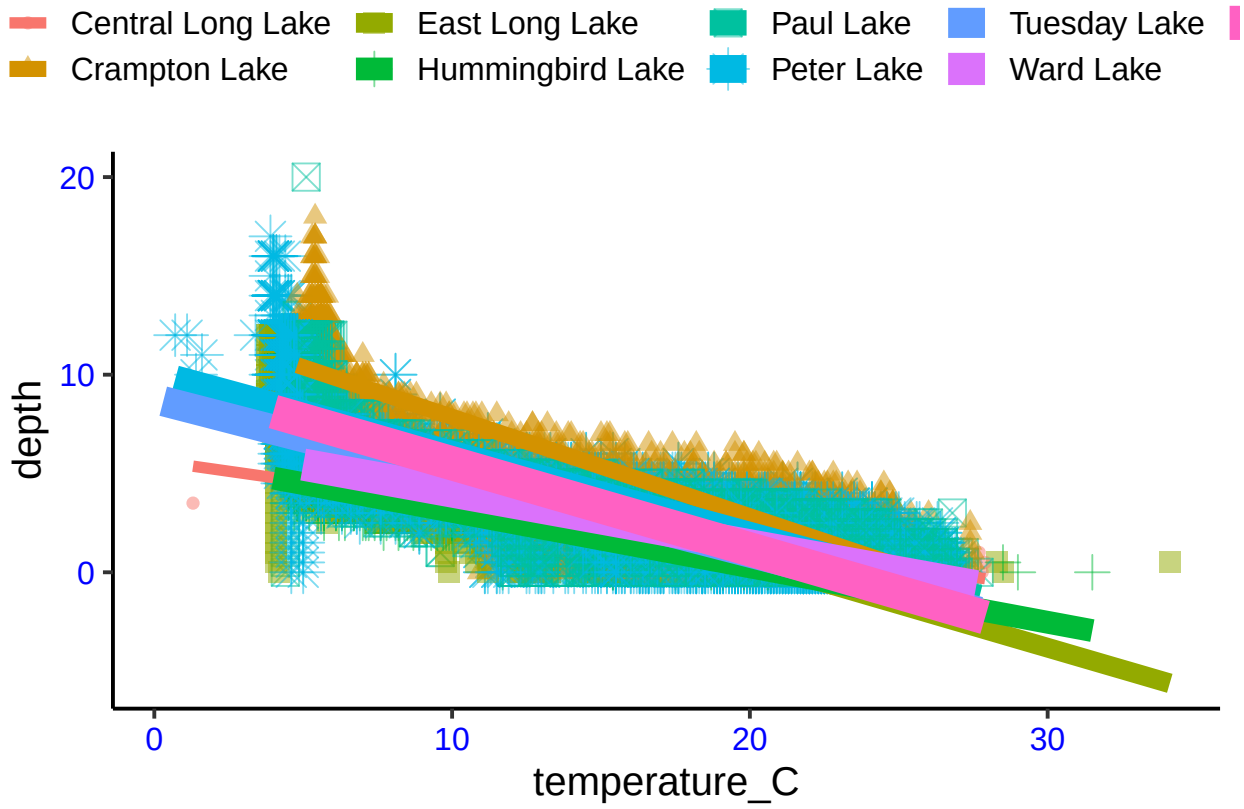
14. Create a graph that depicts temperature by depth, with a separate color for each lake. Add a geom_smooth (method = "lm", se = FALSE) for each lake. Make your points 50 % transparent. Adjust your y axis limits to go from 0 to 35 degrees. Clean up your graph to make it pretty.

```
#14.
CPScatter <-
ggplot(ChemPhysics, aes(x= temperature_C, y=depth, shape= lakename, color=lakename, size=lakename)) +
  geom_point(alpha= .50)+
  geom_smooth(method=lm, se= FALSE)
 labs(x = "Temperature", y = "Depth") +
  ylim(0, 35)
```

```
## NULL
```

```
print(CPScatter)
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.

## Warning: Using size for a discrete variable is not advised.

## ‘geom_smooth()‘ using formula = ’y ~ x’

## Warning: Removed 3858 rows containing non-finite values (‘stat_smooth()‘).

## Warning: The shape palette can deal with a maximum of 6 discrete values because
## more than 6 becomes difficult to discriminate; you have 9. Consider
## specifying shapes manually if you must have them.

## Warning: Removed 13693 rows containing missing values (‘geom_point()‘).
```



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

```
#15
str(CP.Totals.anova)
```

```
## List of 14
##  $ coefficients : Named num [1:7] 211.1 -96.9 -62.8 -50.1 -16.3 ...
##   ..- attr(*, "names")= chr [1:7] "(Intercept)" "lakenameEast Long Lake" "lakenameHummingbird Lake" "
##  $ residuals    : Named num [1:74] 9.57 26.67 -36.23 6.17 -3.83 ...
##   ..- attr(*, "names")= chr [1:74] "117" "134" "135" "240" ...
##  $ effects      : Named num [1:74] -1438.6 93.7 37.4 43 -181.8 ...
##   ..- attr(*, "names")= chr [1:74] "(Intercept)" "lakenameEast Long Lake" "lakenameHummingbird Lake"
##  $ rank         : int 7
##  $ fitted.values: Named num [1:74] 114 114 114 148 148 ...
##   ..- attr(*, "names")= chr [1:74] "117" "134" "135" "240" ...
##  $ assign       : int [1:7] 0 1 1 1 1 1 1
##  $ qr           :List of 5
##   ..$ qr   : num [1:74, 1:7] -8.602 0.116 0.116 0.116 0.116 ...
##   .. ..- attr(*, "dimnames")=List of 2
##   .. .. ..$ : chr [1:74] "117" "134" "135" "240" ...
##   .. .. ..$ : chr [1:7] "(Intercept)" "lakenameEast Long Lake" "lakenameHummingbird Lake" "lakenameP
##   .. ..- attr(*, "assign")= int [1:7] 0 1 1 1 1 1 1
##   .. ..- attr(*, "contrasts")=List of 1
##   .. .. ..$ lakename: chr "contr.treatment"
##   ..$ qraux: num [1:7] 1.12 1.51 1 1 1 ...
```

```
##   ..$ pivot: int [1:7] 1 2 3 4 5 6 7
##   ..$ tol  : num 1e-07
##   ..$ rank : int 7
##   ..- attr(*, "class")= chr "qr"
##  $ df.residual  : int 67
##  $ na.action    : 'omit' Named int [1:1828] 1 2 3 4 5 6 7 8 9 10 ...
##   ..- attr(*, "names")= chr [1:1828] "1" "2" "3" "4" ...
##  $ contrasts    :List of 1
##   ..$ lakename: chr "contr.treatment"
##  $ xlevels      :List of 1
##   ..$ lakename: chr [1:7] "Crampton Lake" "East Long Lake" "Hummingbird Lake" "Paul Lake" ...
##  $ call         : language aov(formula = temperature_C ~ lakename, data = ChemPhysics.july)
##  $ terms        :Classes 'terms', 'formula'  language temperature_C ~ lakename
##   .. ..- attr(*, "variables")= language list(temperature_C, lakename)
##   .. ..- attr(*, "factors")= int [1:2, 1] 0 1
##   .. .. ..- attr(*, "dimnames")=List of 2
##   .. .. .. ..$ : chr [1:2] "temperature_C" "lakename"
##   .. .. .. ..$ : chr "lakename"
##   .. ..- attr(*, "term.labels")= chr "lakename"
##   .. ..- attr(*, "order")= int 1
##   .. ..- attr(*, "intercept")= int 1
##   .. ..- attr(*, "response")= int 1
##   .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
##   .. ..- attr(*, "predvars")= language list(temperature_C, lakename)
##   .. ..- attr(*, "dataClasses")= Named chr [1:2] "numeric" "factor"
##   .. .. ..- attr(*, "names")= chr [1:2] "temperature_C" "lakename"
##  $ model        :'data.frame':   74 obs. of  2 variables:
##   ..$ temperature_C: num [1:74] 124 141 78 154 144 ...
##   ..$ lakename     : Factor w/ 7 levels "Crampton Lake",..: 2 2 2 3 3 3 4 4 4 4 ...
##   ..- attr(*, "terms")=Classes 'terms', 'formula'  language temperature_C ~ lakename
##   .. .. ..- attr(*, "variables")= language list(temperature_C, lakename)
##   .. .. ..- attr(*, "factors")= int [1:2, 1] 0 1
##   .. .. .. ..- attr(*, "dimnames")=List of 2
##   .. .. .. .. ..$ : chr [1:2] "temperature_C" "lakename"
##   .. .. .. .. ..$ : chr "lakename"
##   .. .. ..- attr(*, "term.labels")= chr "lakename"
##   .. .. ..- attr(*, "order")= int 1
##   .. .. ..- attr(*, "intercept")= int 1
##   .. .. ..- attr(*, "response")= int 1
##   .. .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
##   .. .. ..- attr(*, "predvars")= language list(temperature_C, lakename)
##   .. .. ..- attr(*, "dataClasses")= Named chr [1:2] "numeric" "factor"
##   .. .. .. ..- attr(*, "names")= chr [1:2] "temperature_C" "lakename"
##   ..- attr(*, "na.action")= 'omit' Named int [1:1828] 1 2 3 4 5 6 7 8 9 10 ...
##   .. ..- attr(*, "names")= chr [1:1828] "1" "2" "3" "4" ...
##  - attr(*, "class")= chr [1:2] "aov" "lm"
```

```
TukeyHSD(CP.Totals.anova)
```

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

```
## $lakename
##                                       diff        lwr        upr     p adj
## East Long Lake-Crampton Lake    -96.916667 -248.71410  54.880767 0.4614693
## Hummingbird Lake-Crampton Lake  -62.816667 -214.61410  88.980767 0.8682843
## Paul Lake-Crampton Lake         -50.150000 -174.86432  74.564318 0.8830144
## Peter Lake-Crampton Lake        -16.266667 -137.70461 105.171280 0.9996218
## Tuesday Lake-Crampton Lake      -73.055556 -196.99764  50.886530 0.5582194
## West Long Lake-Crampton Lake    -82.700000 -248.98576  83.585757 0.7368665
## Hummingbird Lake-East Long Lake  34.100000 -101.67175 169.871752 0.9876065
## Paul Lake-East Long Lake         46.766667  -57.85249 151.385828 0.8212222
## Peter Lake-East Long Lake        80.650000  -20.04103 181.341026 0.2007073
## Tuesday Lake-East Long Lake      23.861111  -79.83628 127.558500 0.9921992
## West Long Lake-East Long Lake    14.216667 -137.58077 166.014101 0.9999532
## Paul Lake-Hummingbird Lake       12.666667  -91.95249 117.285828 0.9997894
## Peter Lake-Hummingbird Lake      46.550000  -54.14103 147.241026 0.7971086
## Tuesday Lake-Hummingbird Lake   -10.238889 -113.93628  93.458500 0.9999361
## West Long Lake-Hummingbird Lake -19.883333 -171.68077 131.914101 0.9996677
## Peter Lake-Paul Lake             33.883333  -17.59368  85.360348 0.4238795
## Tuesday Lake-Paul Lake          -22.905556  -80.04003  34.228923 0.8844843
## West Long Lake-Paul Lake        -32.550000 -157.26432  92.164318 0.9848660
## Tuesday Lake-Peter Lake         -56.788889 -106.36572  -7.212055 0.0146837
## West Long Lake-Peter Lake       -66.433333 -187.87128  55.004614 0.6425652
## West Long Lake-Tuesday Lake      -9.644444 -133.58653 114.297641 0.9999843
```

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: West Long Lake and Tuesday lake have the same mean temperature as peter lake. Meanwhile East Long Lake and Crapton Lake are statistically distinct from all the other lakes.

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 look further into using the HSD.test function.

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?

```
ChemPhysics.lake <- ChemPhysics %>%
  filter(lakename == "Crampton Lake") %>%
  select(lakename,year4,sampledate,temperature_C) %>%
  drop_na()

lake.anova.2way <- aov(data = ChemPhysics.lake, temperature_C ~  sampledate)
summary(lake.anova.2way)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## sampledate     1     14   14.01   0.303  0.582
## Residuals   1106  51199   46.29
```

Answer: The means temperatures for the lake are not equal in this case and dont match my answer in part 16.