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

Suad Muradov

OVERVIEW

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

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Fay_A06_GLMs.Rmd") prior to submission.

The completed exercise is due on Monday, February 28 at 7:00 pm.

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
getwd()
```

```
## [1] "/Users/smuradov2021/Desktop/Desktop/MY DOCUMENTS/iMEP/Semester 4/ENV872/Environmental_Data_Analy
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

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

```
library(agricolae)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

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

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

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

```
NTLchemphys <-read.csv("./Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv",stringsAsFactors = TRUE)

# Set date to date format
NTLchemphys$sampledate <- as.Date( NTLchemphys$sampledate, format = "%m/%d/%y")

#2 creating own theme
own.theme<-theme_classic(base_size = 18) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "right", legend.title = element_text(size = 14))
theme_set(own.theme)
```

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 temperature recorded during July has strong correlation with depth variable Ha: Mean temperature recorded during July does not have strong correlation with depth variable
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)

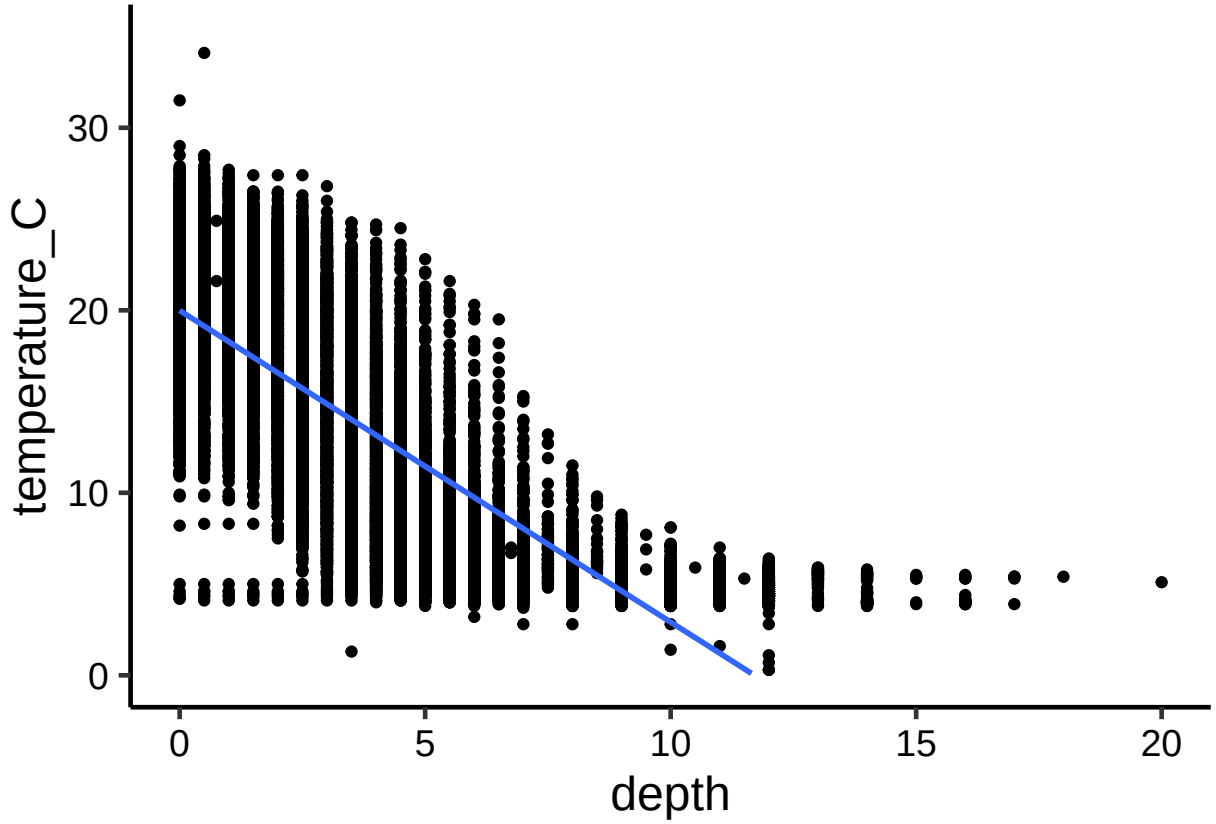
5. Visualize the relationship among the two continuous variables with a scatter plot of temperature by depth. Add a smoothed line showing the linear model, and limit temperature values from 0 to 35 °C. Make this plot look pretty and easy to read.

```
#4
NTLchemphys.processed<-
  NTLchemphys%>%
  select(lakename:daynum, depth, temperature_C)%>%
  na.omit()

#5
NTL.tempbydepth<-
  ggplot(NTLchemphys.processed, aes(x=depth, y=temperature_C), col=)+
  ylim(0,35)+
  geom_point()+
  geom_smooth(method=lm)
print(NTL.tempbydepth)
```

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

```
## Warning: Removed 33 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 scatter plot with smoothed linear model shows clearly that we have a donwward-sloping curve, signifying the negative relationship between the depth and temperature: as we dive deeper, the temperature decreases. The relationship is linear.

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

```
#7

Tempbydepth.regression<-lm(data = NTLchemphys.processed, depth~temperature_C)
summary(Tempbydepth.regression)
```

```
##
## Call:
## lm(formula = depth ~ temperature_C, data = NTLchemphys.processed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7951 -1.3078 -0.2006  1.1548 12.5604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.453719   0.020170   468.7   <2e-16 ***
## temperature_C -0.394920   0.001473  -268.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.905 on 34754 degrees of freedom
## 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: We look at R-squared to measure the level of variability in the constructed model. In our case, both multiple and adjusted R-squared values attest to the fact that *67,42%* of variability being captured by the proposed model and *df is 34754*. As our p-value is smaller than 0,05 the estimates are statistically significant. Model also predicts that the temperature will decrease by *0,395 degrees Celcius* per m of increase in depth.

Multiple regression

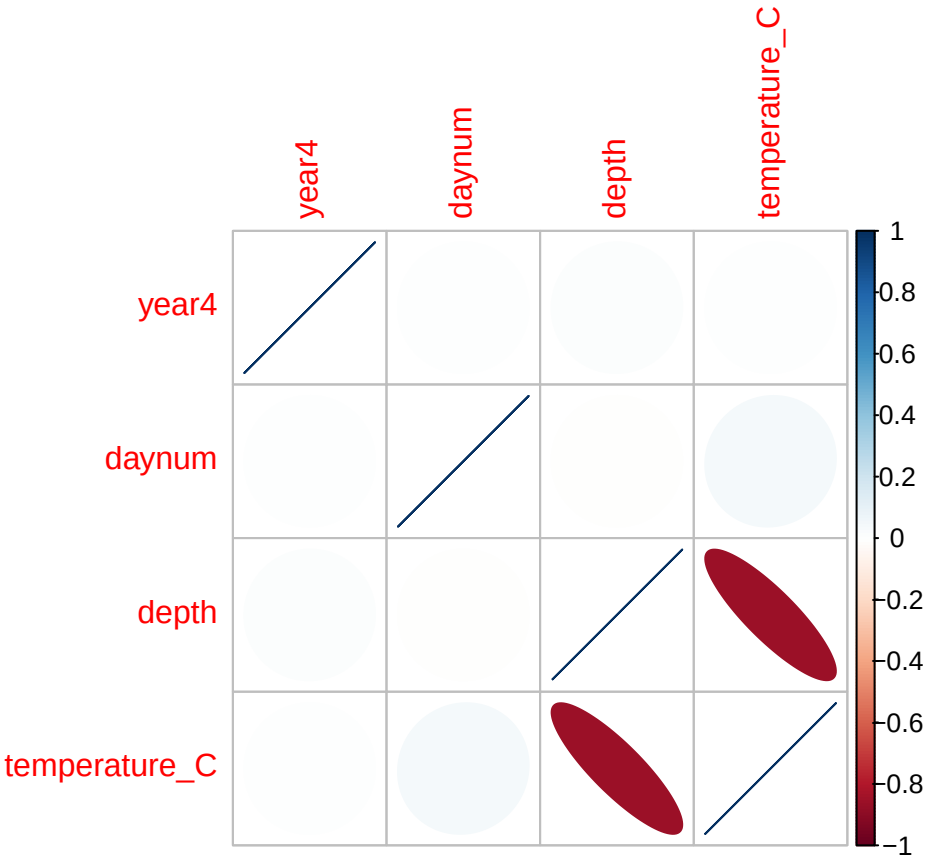
Let's tackle a similar question from a different approach. Here, we want to explore what might the best set of predictors for lake temperature in July across the monitoring period at the North Temperate Lakes LTER.

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

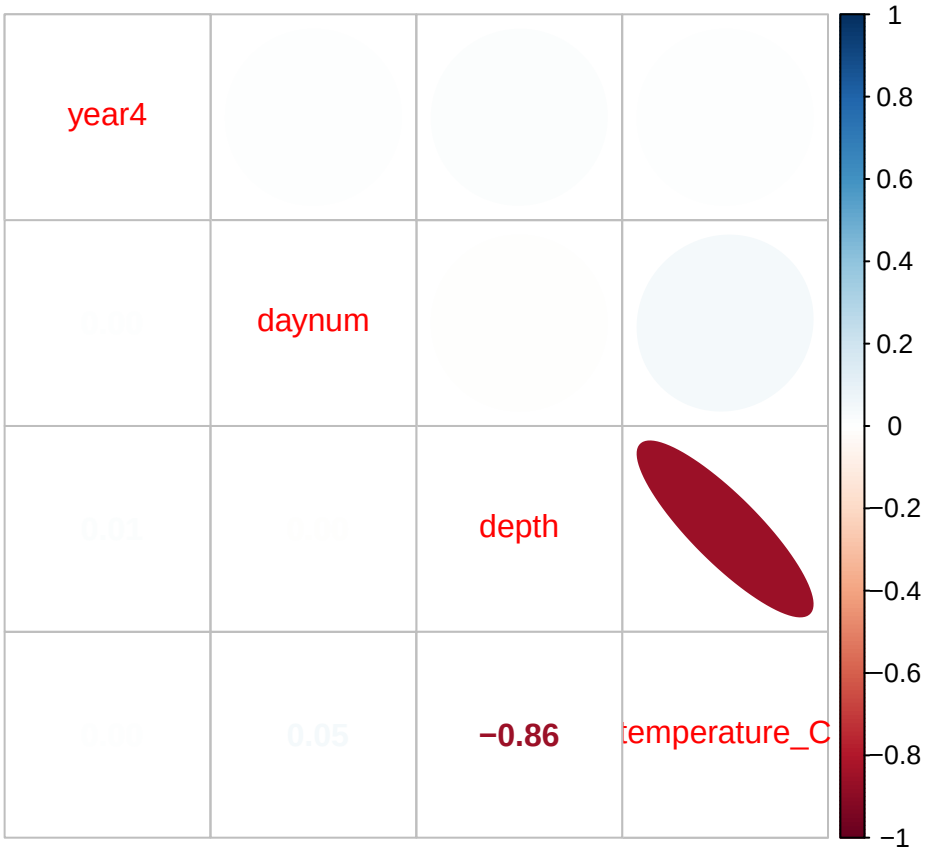
10. Run a multiple regression on the recommended set of variables.

```
#9
NTLchemphys.processed.naomit<-
  NTLchemphys%>%
  filter(month(sampledate)==7)%>%
  select(year4, daynum, depth, temperature_C)%>%
  na.omit

NTLchemphys.processed.cor<-cor(NTLchemphys.processed.naomit)
corrplot(NTLchemphys.processed.cor, method="ellipse")
```




```
corrplot.mixed(NTLchemphys.processed.cor,upper ="ellipse" )
```




```
NTLchemphys.AIC<-lm(data = NTLchemphys.processed.naomit, temperature_C ~ year4 + daynum + depth)
step(NTLchemphys.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 = NTLchemphys.processed.naomit)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##    -8.57556      0.01134      0.03978     -1.94644
```

```
#10

NTLchemphys.new.model<-lm(data = NTLchemphys.processed.naomit, temperature_C ~ year4 + daynum + depth)
summary(NTLchemphys.new.model)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTLchemphys.processed.naomit)
```

```
##
## 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: We were limited by the number of maximum 3 independent variables, and interestingly, the final set included all of them: year4, daynum, depth. New model explains *74,12%* of total variabilities which means it is now *better by the previous model by more than 6%*.

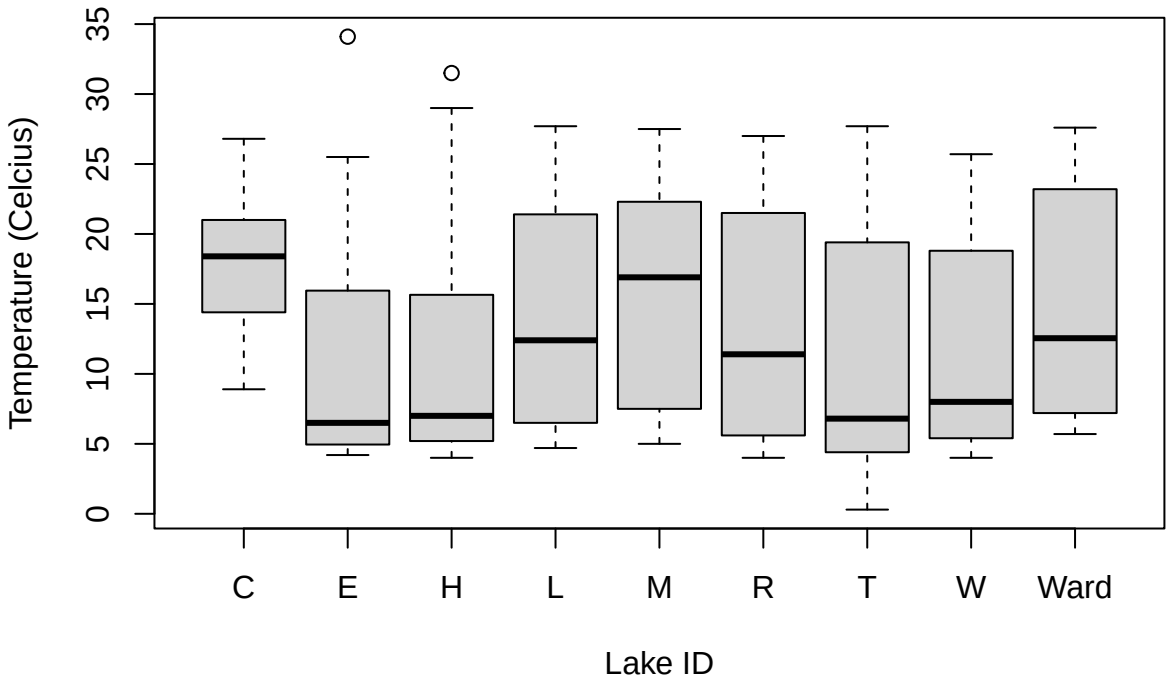
Analysis of Variance

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
# H0: There is no difference between mean temperature as we move through different lakes during July
# H1: Not all lakes have same mean temperatures during July

NTLchemphys.Temp<-NTLchemphys%>%
  select(lakeid, lakename, sampledate, temperature_C)%>%
  filter(month(sampledate)==7)%>%
  na.omit()

boxplot(NTLchemphys.Temp$temperature_C~NTLchemphys.Temp$lakeid, ylab = "Temperature (Celcius)", xlab =
```




```
NTLchemphys.Temp.anova <- aov(data = NTLchemphys.Temp, temperature_C~lakeid)
summary(NTLchemphys.Temp.anova)
```

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

```
NTLchemphys.Temp.anova2 <- lm(data = NTLchemphys.Temp, temperature_C~lakeid)
summary(NTLchemphys.Temp.anova2)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakeid, data = NTLchemphys.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 ***
## lakeidE      -7.3987     0.6918 -10.695  < 2e-16 ***
## lakeidH      -6.8931     0.9429  -7.311 2.87e-13 ***
## lakeidL      -3.8522     0.6656  -5.788 7.36e-09 ***
## lakeidM      -2.3145     0.7699  -3.006 0.002653 **
## lakeidR      -4.3501     0.6645  -6.547 6.17e-11 ***
## lakeidT      -6.5972     0.6769  -9.746  < 2e-16 ***
## lakeidW      -6.0878     0.6895  -8.829  < 2e-16 ***
## lakeidWard   -3.2078     0.9429  -3.402 0.000672 ***
## ---
## 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
```

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

    Answer: As we started, our null hypothesis was that all means of all lakes have the same mean temperature in July. As we first performed a boxplot analysis and then conducted anova in two models, we found out that temperature was very significant with 3 stars (p-value<0.05) to lakeid which meant that there is a difference between lakes when it comes to the average July temperatures.

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.

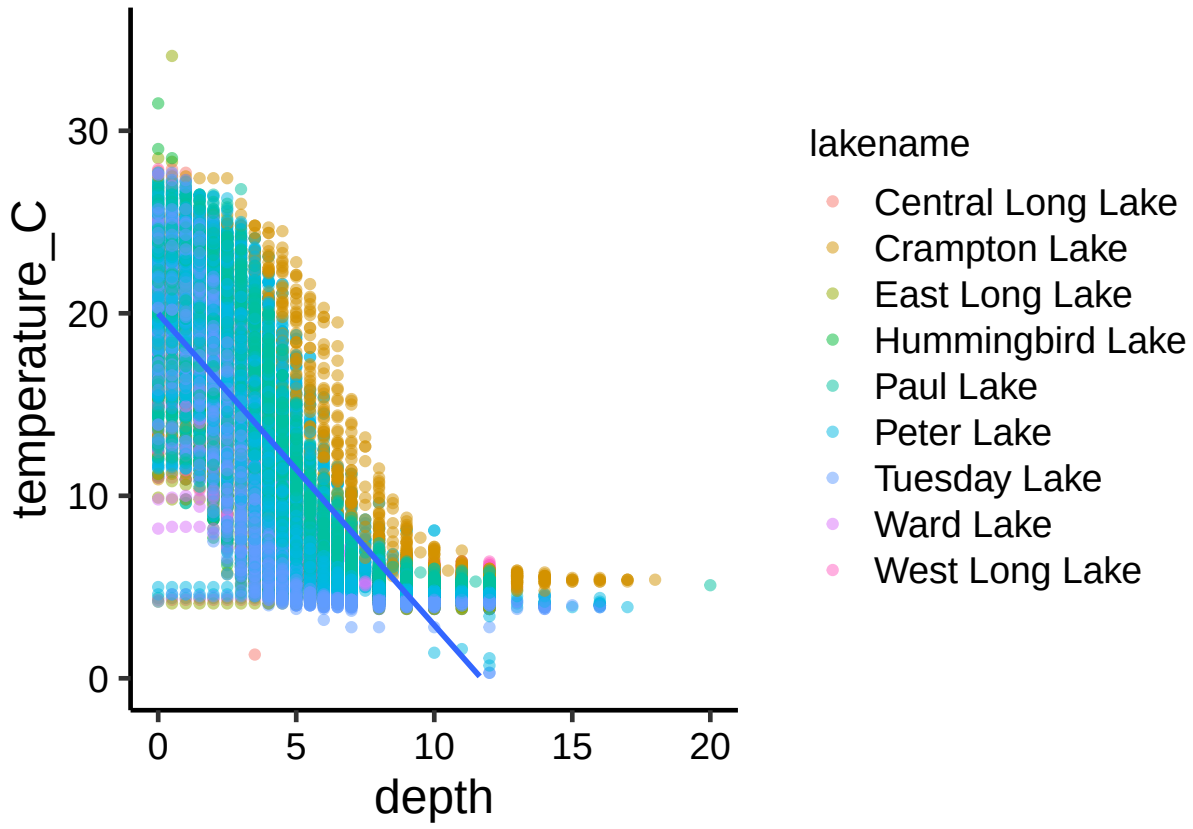
```
#14.
NTLchemphys.TempbyDep<-
  ggplot(NTLchemphys, aes(x = depth, y = temperature_C))+
  ylim(0,35)+
  geom_point(aes(color = lakename), alpha=0.5)+
  geom_smooth(method = "lm", se = FALSE)
print(NTLchemphys.TempbyDep)
```

```
## `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).
```

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



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

```
#15
TukeyHSD(NTLchemphys.Temp.anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakeid, data = NTLchemphys.Temp)
##
## $lakeid
##               diff        lwr        upr     p adj
## E-C    -7.3987410 -9.5449411 -5.2525408 0.0000000
## H-C    -6.8931304 -9.8184178 -3.9678430 0.0000000
## L-C    -3.8521506 -5.9170942 -1.7872070 0.0000003
## M-C    -2.3145195 -4.7031913  0.0741524 0.0661566
## R-C    -4.3501458 -6.4115874 -2.2887042 0.0000000
## T-C    -6.5971805 -8.6971605 -4.4972005 0.0000000
## W-C    -6.0877513 -8.2268550 -3.9486475 0.0000000
## Ward-C -3.2077856 -6.1330730 -0.2824982 0.0193405
## H-E     0.5056106 -1.7364925  2.7477137 0.9988050
## L-E     3.5465903  2.6900206  4.4031601 0.0000000
## M-E     5.0842215  3.6092730  6.5591700 0.0000000
## R-E     3.0485952  2.2005025  3.8966879 0.0000000
## T-E     0.8015604 -0.1363286  1.7394495 0.1657485
## W-E     1.3109897  0.2885003  2.3334791 0.0022805
## Ward-E  4.1909554  1.9488523  6.4330585 0.0000002
## L-H     3.0409798  0.8765299  5.2054296 0.0004495
```

```
## M-H     4.5786109  2.1034131  7.0538088 0.0000004
## R-H     2.5429846  0.3818755  4.7040937 0.0080666
## T-H     0.2959499 -1.9019508  2.4938505 0.9999752
## W-H     0.8053791 -1.4299320  3.0406903 0.9717297
## Ward-H  3.6853448  0.6889874  6.6817022 0.0043297
## M-L     1.5376312  0.1836408  2.8916215 0.0127491
## R-L    -0.4979952 -1.1120620  0.1160717 0.2241586
## T-L    -2.7450299 -3.4781416 -2.0119182 0.0000000
## W-L    -2.2356007 -3.0742314 -1.3969699 0.0000000
## Ward-L  0.6443651 -1.5200848  2.8088149 0.9916978
## R-M    -2.0356263 -3.3842699 -0.6869828 0.0000999
## T-M    -4.2826611 -5.6895065 -2.8758157 0.0000000
## W-M    -3.7732318 -5.2378351 -2.3086285 0.0000000
## Ward-M -0.8932661 -3.3684639  1.5819317 0.9714459
## T-R    -2.2470347 -2.9702236 -1.5238458 0.0000000
## W-R    -1.7376055 -2.5675759 -0.9076350 0.0000000
## Ward-R  1.1423602 -1.0187489  3.3034693 0.7827037
## W-T     0.5094292 -0.4121051  1.4309636 0.7374387
## Ward-T  3.3893950  1.1914943  5.5872956 0.0000609
## Ward-W  2.8799657  0.6446546  5.1152769 0.0021080
```

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: Ward Lake is the closest to Peter Lake based on the p-value of 0.78 which is the highest p-value among bilateral comparisons of Peter Lake. Central Long Lake is statistically distinct from all the other lakes, especially when the significance level is set at 0.1

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: Since we need to test just Peter Lake and Paul Lake, two variables, regarding them having distinct mean temperatures, we could use two-way ANOVA. It allows us to examine the effects of two categorical explanatory variables on a continuous response variable and thus, would work perfectly here.