

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

David Amanfu

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.

```
getwd()
```

```
## [1] "/Users/davidamanfu/Desktop/Duke MPP/Environ Data /Environmental_Data_Analytics_2022/Assignments"
```

```
knitr::opts_knit$set(root.dir = "~/Desktop/Duke MPP/Environ Data /Environmental_Data_Analytics_2022/")
```

```
#1  
library(tidyverse)
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```

library(agricolae)
library(lubridate)
library(cowplot)
library(corrplot)

NTL_LTER <- read.csv("./Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv", stringsAsFactors = TRUE)
NTL_LTER$sampldate <- as.Date(NTL_LTER$sampldate, format = "%m/%d/%y")

#2
library(ggthemes)
library(hrbrthemes)
library(extrafont)
library(extrafontdb)
AmanfuTheme2 <- theme_ipsum()+
  theme(legend.position = "bottom",
        legend.key = element_rect(fill = "white", colour = "black"), legend.direction = "horizontal",
        legend.title = element_text(face = "bold"))
theme_set(AmanfuTheme2)

```

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: For all lakes with temperature recordings in July, the lake temperatures do not vary with depth. There is no correlation between lake temperature and depth. Ha: For all lakes with temperature recordings in July, the lake temperatures vary with depth. There is a correlation between lake temperature and depth.
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: `lakenname`, `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.

I'll note it's a bit unclear whether you're asking for a linear regression or just a smoothed curve fit. I'm including both.

```

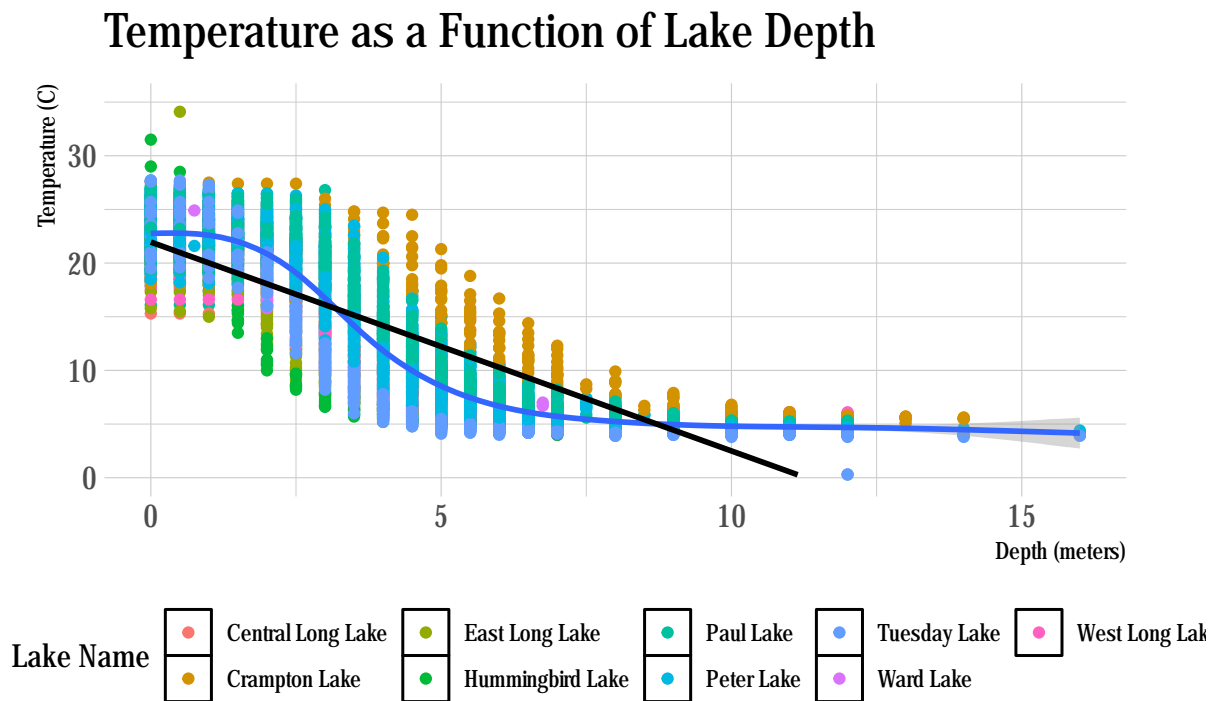
#4
NTL_LTER_P <- NTL_LTER %>%
  filter(month(sampldate) == 7) %>%
  select(lakenname, year4, daynum, depth, temperature_C) %>%
  #drop_na() %>%
  na.omit()
#summary(NTL_LTER_P)

```

```
#5
LTER_plot <- ggplot(NTL_LTER_P, aes(x=depth, y=temperature_C)) +
  geom_point(aes(color=lake_name), show.legend = TRUE) +
  ylim(0,35)+
  labs(color = "Lake Name", title = "Temperature as a Function of Lake Depth", x = "Depth (meters)", y=
  geom_smooth(na.rm= TRUE) +
  geom_smooth(method=lm,na.rm=TRUE,color="black",se = TRUE)
print(LTER_plot)
```

```
## 'geom_smooth()' using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

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



6. Interpret the figure. What does it suggest with regards to the response of temperature to depth? Do the distribution of points suggest anything about the linearity of this trend?

Answer: Taking the linear fit $y = \alpha * x + \beta$ plotted in black on the figure we can assess that the lake gets colder the further you go down. It's not a perfect fit but certainly it tracks.

The distribution of the points tells a more particular story though. If we refer to the `geom_smooth`, we can see that it follows a sigmoid-looking function. At shallow depths, the temperature holds near constant, similar to the surface temperature, but after a threshold of one (1) meter in depth, the temperature begins to drop. After about 8 meters it levels off again.

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

```

#7
depth_regr <- lm(data= NTL_LTER_P, temperature_C ~ depth)
summary(depth_regr)

##
## Call:
## lm(formula = temperature_C ~ depth, data = NTL_LTER_P)
##
## 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 ***
## 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

cor.test(NTL_LTER_P$temperature_C,NTL_LTER_P$depth)

##
## Pearson's product-moment correlation
##
## data:  NTL_LTER_P$temperature_C and NTL_LTER_P$depth
## t = -165.83, df = 9726, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8646036 -0.8542169
## sample estimates:
##          cor
## -0.8594989

```

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:

So what we find is that our linear regression is estimating a function that looks like our black line on the plot above, and has the characteristics/equation of the following: $y = -1.94 \pm 0.01 * x + 21.95 \pm 0.06$

This means for each unit of depth, the temperature decreases (-) by about $2^{\circ}C$, starting at a value of about $22^{\circ}C$ at the surface. We find a high level of statistical significance, and with that are able to reject our null hypothesis. As well, with over 9700 degrees of freedom, we have pretty robust information to calculate our estimates. When we look at our R-squared value, we interpret that about 74% of our variability in temperature can be explained by looking at the corresponding depth where the measurement was taken.

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

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

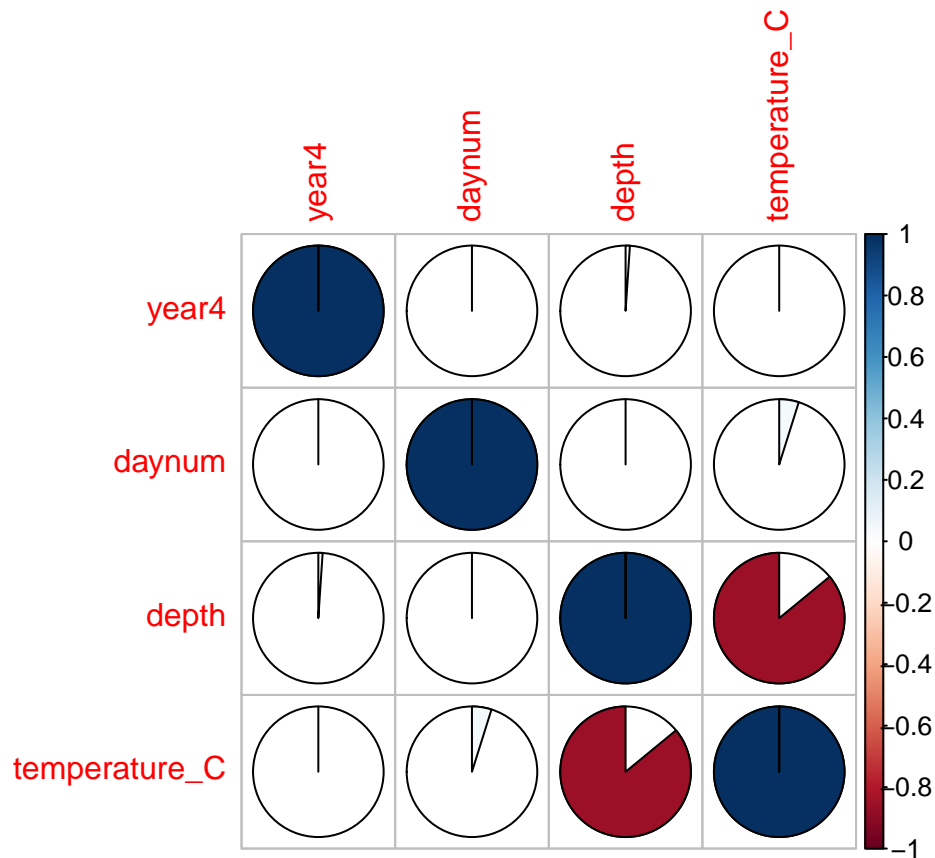
```
step(depth_regr_AIC)
```

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

#10

```
corrplot(cor(NTL_LTER_P[,select(year4:temperature_C)]),method="pie",outline=TRUE)
```



```
step_AIC_regr <- lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL_LTER_P)
summary(step_AIC_regr)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL_LTER_P)
##
## 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 AIC stepwise method recommended using three variables, depth, day number (as in calendar date 1-365; limited to July), and year the sample was collected. This new model improves our previous R-squared mark of 0.7387, and gives us 0.7412, a very modest improvement. It appears that there is some additional correlation between temperature and day of the year, but the correlation is very light. Using the `corrplot` package we can see these relationships pretty quickly in a visual form.

For the question of “is this better” or an improvement, given that there’s statistical significance in adding those extra variables, technically yes. But for the purposes of getting a pretty robust analysis in the first place, the addition of extra variables does not substantially improve our model with some significant level of clarity or insight. Our small inferences suggest that maybe there is a seasonal trend, where it gets hotter as the month of July progresses. But that’s a bit of a trivial thing to say out loud.

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

```
NTL_anova <- aov(data=NTL_LTER_P, temperature_C ~ lakename)
summary(NTL_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
```

```
NTL_lm <- lm(data=NTL_LTER_P, temperature_C ~ lakename)
summary(NTL_lm)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = NTL_LTER_P)
##
```

```
## 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 ***
## lakenamCrampton Lake    -2.3145     0.7699  -3.006 0.002653 **
## lakenamEast Long Lake   -7.3987     0.6918 -10.695 < 2e-16 ***
## lakenamHummingbird Lake -6.8931     0.9429  -7.311 2.87e-13 ***
## lakenamPaul Lake        -3.8522     0.6656  -5.788 7.36e-09 ***
## lakenamPeter Lake       -4.3501     0.6645  -6.547 6.17e-11 ***
## lakenamTuesday Lake    -6.5972     0.6769  -9.746 < 2e-16 ***
## lakenamWard Lake        -3.2078     0.9429  -3.402 0.000672 ***
## lakenamWest 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
```

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

Answer:

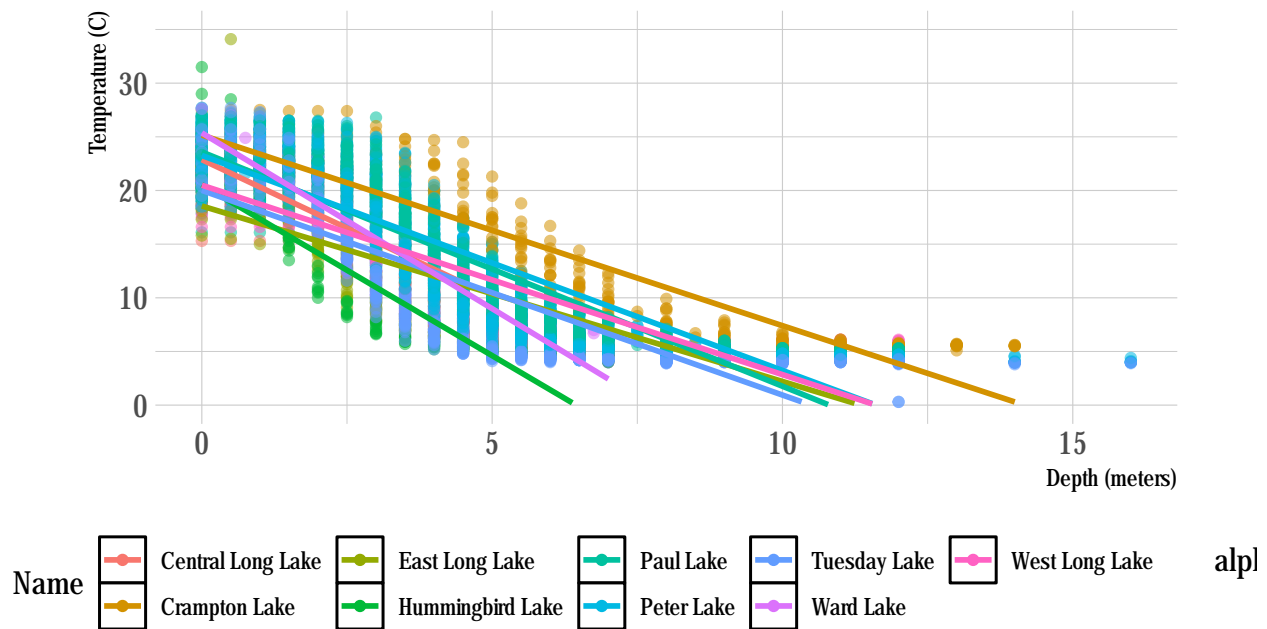
Statistically yes! We see that from the ANOVA test, we

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.
LTER_plot2 <- ggplot(NTL_LTER_P, aes(x=depth, y=temperature_C)) +
  geom_point(aes(color=lakenam, alpha=0.5), show.legend = TRUE) +
  ylim(0,35)+
  labs(color="Lake Name", title = "Temperature as a Function of Lake Depth", x = "Depth (meters)", y=
  geom_smooth(method=lm, na.rm=TRUE, aes(color=lakenam), se = FALSE)
print(LTER_plot2)
```

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


Temperature as a Function of Lake Depth



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

#15

TukeyHSD(NTL_anova)

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

```
#plot(TukeyHSD(NTL_anova))
```

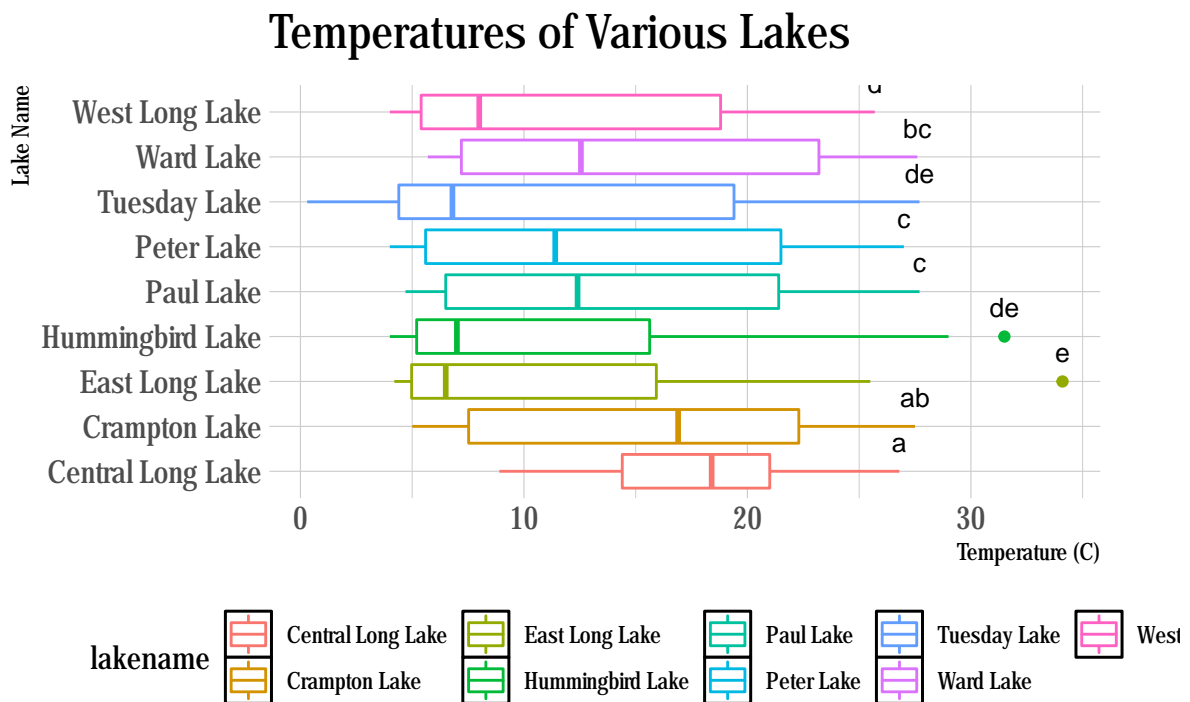
```
NTL_grouped <- HSD.test(NTL_anova,"lakename",group=TRUE)
NTL_grouped
```

```
## $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"
```

```
#class(NTL_grouped$groups)
```

```
NTL_box <- ggplot(NTL_LTER_P,aes(y=temperature_C, x = lakename))+
  geom_boxplot(aes(color=lakename))+
  labs(title="Temperatures of Various Lakes", x="Lake Name", y = "Temperature (C)")+
  coord_flip()+
  stat_summary(geom = "text", fun = max, vjust = -1, size = 3.5, label=c("a","ab","e","de","c","c","de"))
print(NTL_box)
```



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: If we peek at the “NTL_box” boxplot, we see Paul Lake and Ward Lake are grouped

together with Peter Lake, and the boxplot bears this out. The labels have been assigned from the the HSD test's groups, *NTL_grouped\$groups\$groups*, just reordered.

There are no lakes that are statistically distinct from any other lake. Even Central Long Lake, with the highest average temperature, is still in the same statistical league as Crampton 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:

If we were just looking at these two we might look at a humble ttest. It would tell us quickly and efficiently what the differences between means would be, and whether that difference was statistically significant or not.