

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

Ina Liao

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.

```
#install.packages("tidyverse")
#install.packages("agricolae")
#install.packages("here")
#install.packages("ggplot2")
#install.packages("cowplot")
#install.packages("ggridges")
#install.packages("viridis")
#install.packages("RColorBrewer")
#install.packages("colormap")
#install.packages("ggthemes")
#install.packages("dplyr")

library(tidyverse)
library(agricolae)
library(here)
library(ggplot2)
library(cowplot)
library(ggridges)
library(viridis)
library(RColorBrewer)
library(colormap)
library(ggthemes)
library(dplyr)
here()
```

```
df_NTL_LTER<-read.csv(here("Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv"))
#date format
df_NTL_LTER$sampleddate<-as.Date(df_NTL_LTER$sampleddate,format="%m/%d/%y")
```

2. Build a ggplot theme and set it as your default theme.

```
plot_theme<-theme_calc()+
  theme(
    #plot background
    plot.background=element_rect(color="gray"),

    #plot title
    plot.title=element_text(color="black",hjust=0.5,vjust=1),

    #axis labels
    axis.text=element_text(color="black"),

    #legend
    legend.key=element_rect(color="white"),
    legend.background = element_rect(color="white"),
    legend.position="right",

    #gridlines
    panel.grid.major=element_line("white"),
    axis.ticks=element_line(color="white")
  )
theme_set(plot_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: The mean lake temperature recorded during July does not change with depth across all lakes. Ha: The mean lake temperature recorded during July does 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)

```
df_July<-df_NTL_LTER %>%
  mutate(month=format(sampleddate,"%B"))%>%
  filter(month=="July") %>%
  select(lakename,year4,daynum,depth,temperature_C) %>%
  na.omit()
```

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.

```
plot_scatter<-ggplot(df_July,aes(x=temperature_C,y=depth,color=temperature_C))+
  geom_point()+
  xlim(0,35)+
  scale_y_reverse()+
```

```
scale_color_gradient(low = "blue", high = "orange")+
geom_smooth(method="lm",color="black")+
labs(x="temperature (C)",y="depth (m)")
print(plot_scatter)
```

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

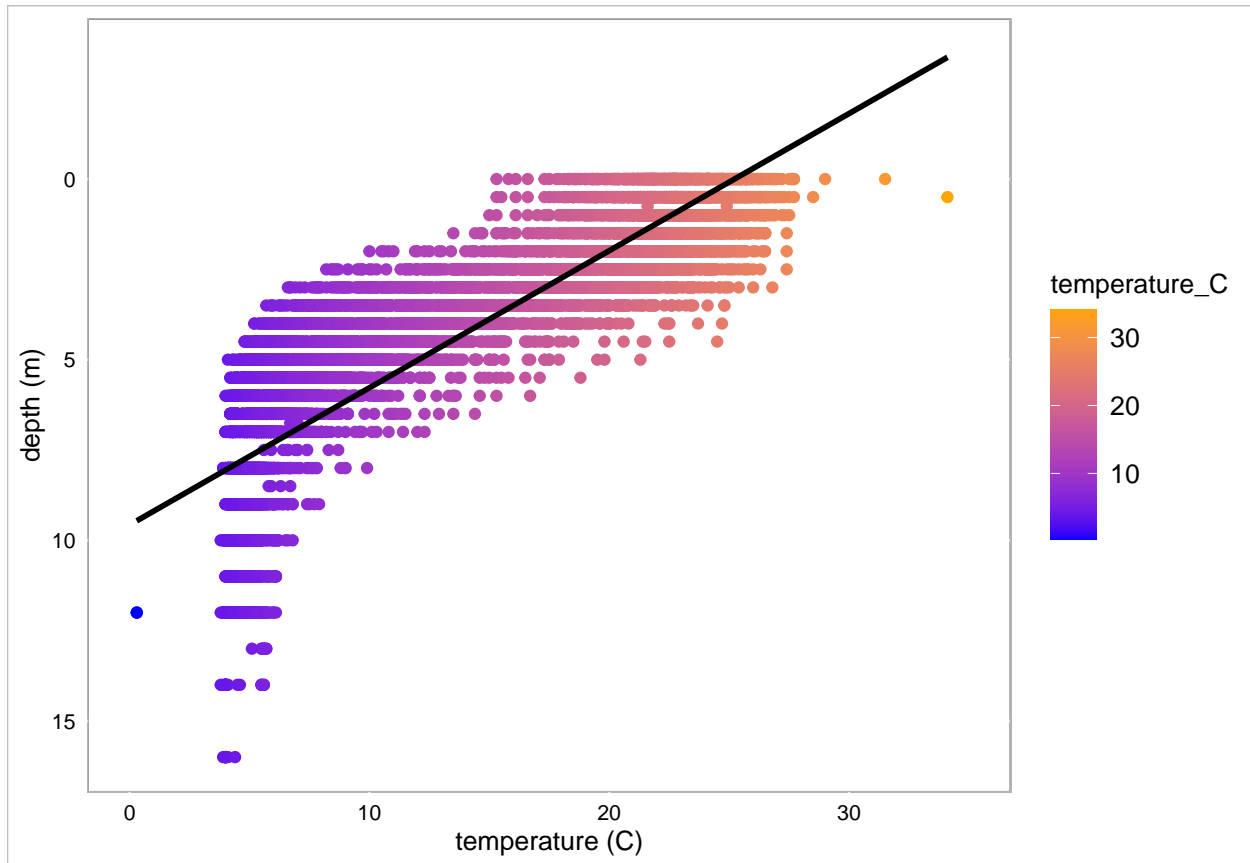


Figure 1: Temperature by depth

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: As the depth of the water increases, the temperature decreases, which means that there might be a negative linear relationship between depth and temperature.

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

```
tem_dep<-lm(df_July$temperature_C~df_July$depth)
summary(tem_dep)
```

```
##
## Call:
## lm(formula = df_July$temperature_C ~ df_July$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 ***
## df_July$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 residual of the model ranges from -9.5 to 13.6. The slope of the regression line is -1.94 (df=9726), meaning there is a negative relationship between temperature and depth level. The coefficient is statistically significantly different from 0 (p-value<0.001), meaning that we have enough evidence to reject that that mean lake temperature recorded during July does not change with depth across all lakes. The r-square value is 0.738, indicating that a unit increase in depth is associated with a 0.738 unit increase in the predicted value of temperature.

Multiple regression

Let's tackle a similar question from a different approach. Here, we want to explore what might be 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.
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?

```
#9 AIC test
test_AIC<-lm(df_July$temperature_C~df_July$year4 + df_July$daynum + df_July$depth)
step(test_AIC)
```

```
## Start:  AIC=26065.53
## df_July$temperature_C ~ df_July$year4 + df_July$daynum + df_July$depth
##
##              Df Sum of Sq    RSS   AIC
## <none>                  141687 26066
## - df_July$year4         1      101 141788 26070
## - df_July$daynum        1     1237 142924 26148
## - df_July$depth         1    404475 546161 39189
##
## Call:
## lm(formula = df_July$temperature_C ~ df_July$year4 + df_July$daynum +
##     df_July$depth)
##
## Coefficients:
## (Intercept)  df_July$year4  df_July$daynum  df_July$depth
##      -8.57556       0.01134       0.03978      -1.94644
```

```
#10 linear model
model_linear<-lm(df_July$temperature_C~ df_July$daynum + df_July$depth)
summary(model_linear)

##
## Call:
## lm(formula = df_July$temperature_C ~ df_July$daynum + df_July$depth)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.6174 -2.9809  0.0845  2.9681 13.4406
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   14.088588   0.855505   16.468  <2e-16 ***
## df_July$daynum  0.039836   0.004318    9.225  <2e-16 ***
## df_July$depth  -1.946111   0.011685  -166.541  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.818 on 9725 degrees of freedom
## Multiple R-squared:  0.741, Adjusted R-squared:  0.741
## F-statistic: 1.391e+04 on 2 and 9725 DF, p-value: < 2.2e-16
```

Answer: The AIC test suggests us to use “daynum” and “depth” as explanatory variables in the multiple regression to predict temperature. The model indicate that a unit increase in daynum and depth is associated with a 0.741 unit increase in the predicted temperature value, which has higher prediction than the previous model (r-square: 0.741>0.728).

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).

Ho: the mean temperatures in each lake is the same in the month of July Ha: the mean temperatures in each lake is not the same in the month of July

```
# wrangle data
df_tem_lake<-df_July %>%
  group_by(lakename,year4,daynum) %>%
  summarise(temperature_C)

## `summarise()` has grouped output by 'lakename', 'year4', 'daynum'. You can
## override using the `.groups` argument.

# one-way ANOVA
tem_ANOVA<-aov(data=df_tem_lake,temperature_C~lakename)
summary(tem_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

# format ANOVA as linear model
tem_ANOVA_lm<-lm(data=df_tem_lake,temperature_C~lakename)
summary(tem_ANOVA_lm)

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

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

Answer: The coefficient is statistically significant different from 0 (p-value<0.001), meaning that we have enough evidence to reject that the temperatures in each lake is the same in the month of July.

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_tem_lake<-ggplot(df_July,aes(x=temperature_C,y=depth,color=lakename))+
  geom_point(alpha=0.5)+
  xlim(0,35)+
  scale_y_reverse()+
  labs(x="temperature (C)",y="depth (m)",color="Lake Name")+
  theme(legend.position="right",
        legend.text = element_text(size = 10),
        legend.title = element_text(size = 10))
print(plot_tem_lake)
```

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

```
#15
# HSD test
TukeyHSD(tem_ANOVA)
```

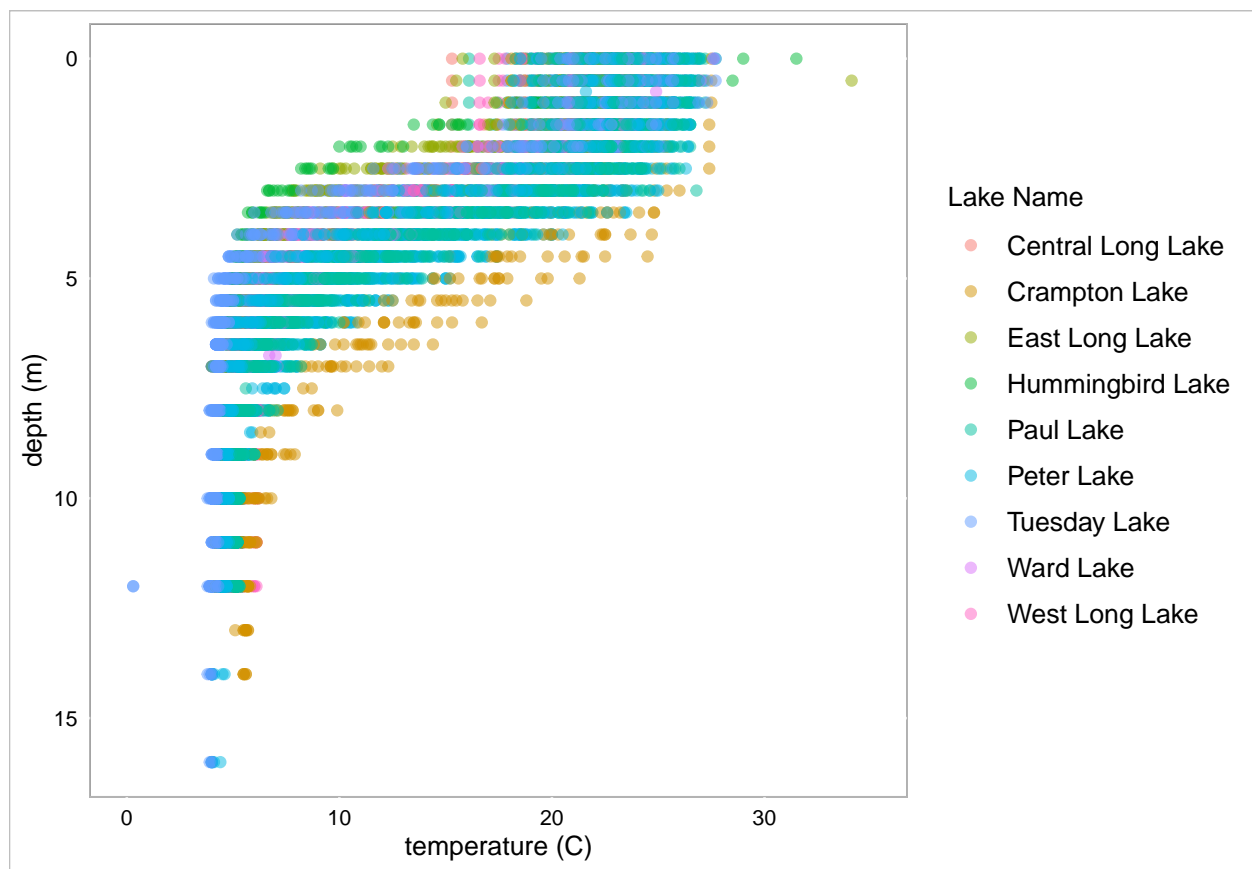


Figure 2: Temperature by depth in each lake

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

```
test_HSD<-HSD.test(tem_ANOVA,"lakename",group = TRUE)
test_HSD
```

```
## $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"
```

```
# graph the results
plot_HSD<-ggplot(df_tem_lake, aes(x = lakename, y = temperature_C)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  stat_summary(geom = "text", fun = max, vjust = -1, size = 3.5,
    label = c("a", "b", "c", "d", "de", "e",
      "e", "f", "f")) +
  labs(x = "Lakename", y = "Temperature (C)")
print(plot_HSD)
```

All lakes are grouped with at least another lake; thus, no lake has a different means than others.

16. From the findings above, which lakes have the same mean temperature, statistically speaking, as Peter

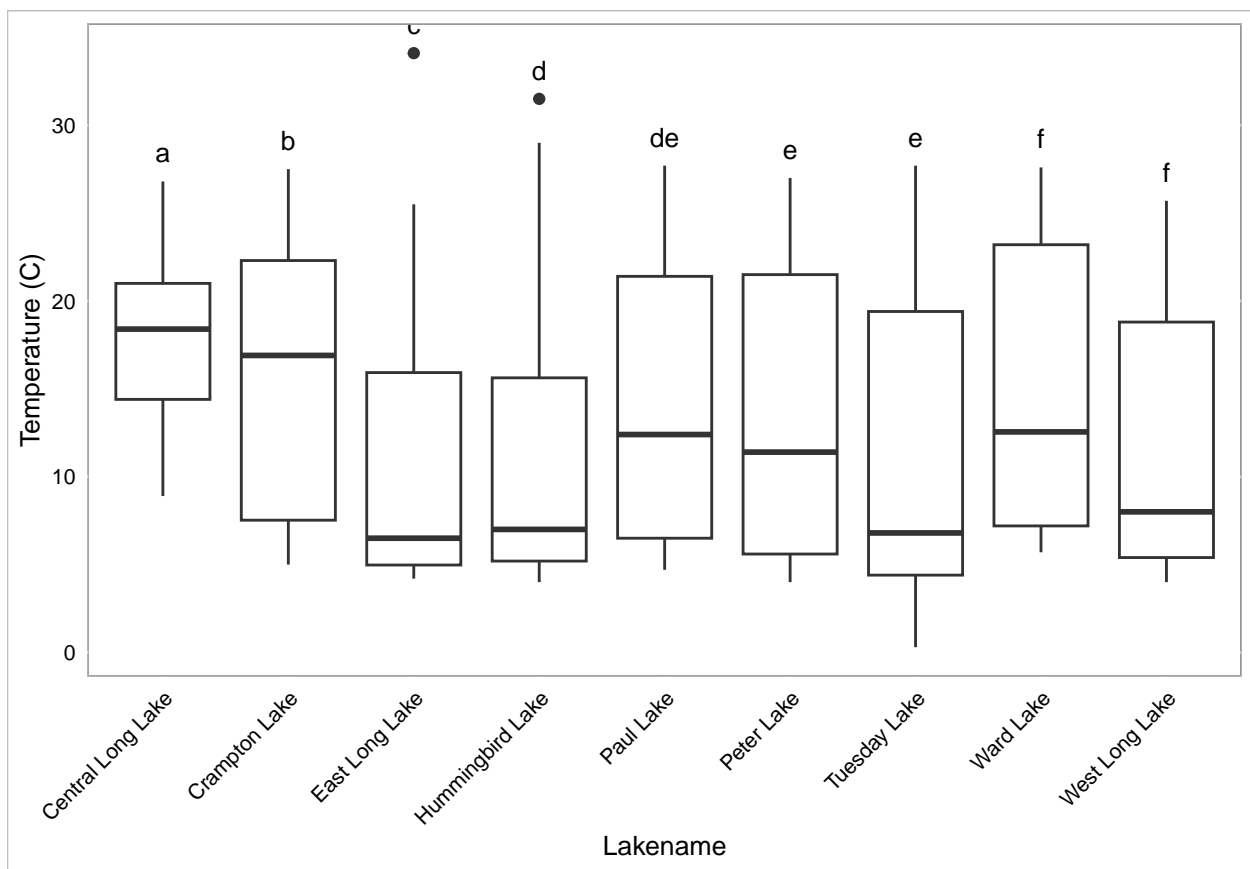


Figure 3: Mean temperature in each lake

Lake? Does any lake have a mean temperature that is statistically distinct from all the other lakes?

Answer: Statistically, Peter Lake has the same mean temperature as Paul Lake ($p > 0.001$).

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 can look into the interaction effect of Peter Lake and Paul Lake on the mean temperature to understand if the effect of Peter Lake temperature and Paul Lake temperature depends on each other.

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 your answer for part 16?

Ho: the mean temperature in Crampton Lake and Ward Lake is the same in the month of July
Ha: the mean temperature in Crampton Lake and Ward Lake is not the same in the month of July

```
# filter data
df_July_sub <- df_July %>%
  filter(lakename == "Crampton Lake" | lakename == "Ward Lake")

# two-sample t-test
test_twosample <- t.test(df_July_sub$temperature_C ~ df_July_sub$lakename)
test_twosample
```

```
##
## Welch Two Sample t-test
##
## data: df_July_sub$temperature_C by df_July_sub$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 not equal to 0
## 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: We do not have enough evidence to reject that the mean temperature in Crampton Lake and Ward Lake are the same ($p\text{-value} > 0.1$, $df = 200$), which matches the answer in question 16.