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

Queenie Wei

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)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.1
```

```
## Warning: package 'lubridate' was built under R version 4.3.1
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- 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
```

```r
options(scipen = 4)
library(here)
```

```
## Warning: package 'here' was built under R version 4.3.1
```

```
## here() starts at C:/Users/ziyaw/Downloads/EDE_Fall2023
```

```r
here()
```

```
## [1] "C:/Users/ziyaw/Downloads/EDE_Fall2023"
```

```r
library(dplyr)
library(agricolae)
```

```
## Warning: package 'agricolae' was built under R version 4.3.1
```

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



#2
# Set theme
mytheme <- theme_classic(base_size = 13) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "right")
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: The mean lake temperature recorded during July doesnt change with depth across all lakes Ha: The mean lake temperature recorded during July changes with depth across all lakes
4. Wrangle your NTL-LTER dataset with a pipe function so that the records meet the following criteria:

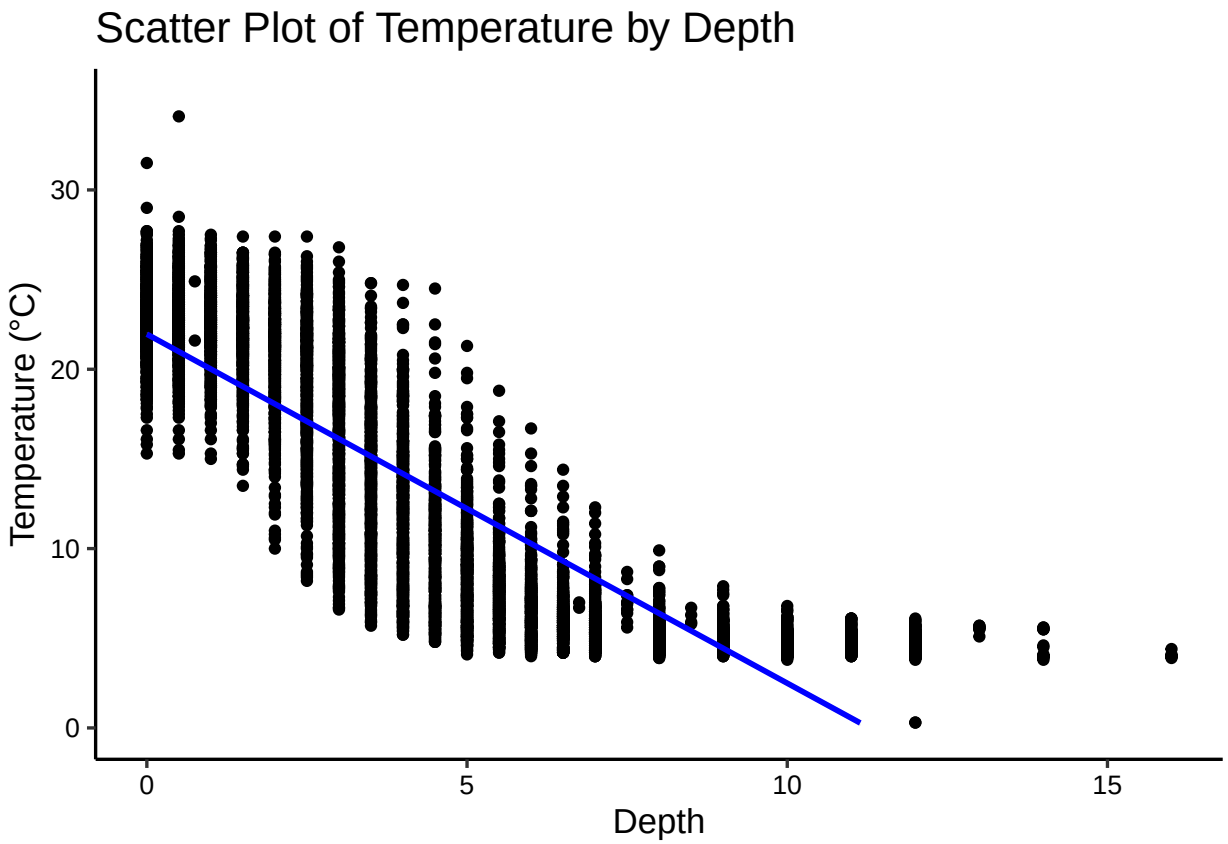
- Only dates in July.
- Only the columns: `lakename`, `year4`, `daynum`, `depth`, `temperature_C`
- Only complete cases (i.e., remove NAs)

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.

```r
#4
### wrangling the dataset so that it only contains: dates in july, only the
### following columns: `lakename`, `year4`, `daynum`, `depth`, `temperature_C`,
### omits NAs
NTL_filtered <- NTL %>%
  filter(format(sampledate, "%m") == "07") %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  na.omit()

#5
###visualizing the dataset using a scatter plot of temp by depth. Adding a
## lm line to it, and limiting the y_axis to 0 to 35
# Create the scatter plot
ggplot(NTL_filtered, aes(x = depth, y = temperature_C)) +
  geom_point() +  # Add scatter points
  geom_smooth(method = "lm", formula = y ~ x, color = "blue") +
  # Add a linear model smoothed line
  labs(x = "Depth", y = "Temperature (°C)") +  # Label the axes
  ggtitle("Scatter Plot of Temperature by Depth") +  # Add a title
  mytheme +
  scale_y_continuous(limits = c(0, 35))
```

```
## Warning: Removed 24 rows containing missing values (‘geom_smooth()‘).
```



```
# Limit temperature values from 0 to 35 °C
###works cited: Chatgpt. I fed chatgpt my code to question 4 and the prompt of
#this question. Then I changed the theme to mytheme
```

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: There is a negative response of temperature to depth. The distribution of points (the fact that most of them are close to the line and also kind of evenly distributed) suggests that there is a strong negative linear relationship between the two variables. ###works cited: Bard. Prompt: I now have a scatter plot of lake temperature by depth with a smooth line showing the linear model. how does the distribution of points suggest in regard to the linearity of this trend? I then asked for code that can give me more information on the linearity of the trend and Bard gave me "#cor(NTL_filtered$temperature_C, NTL_filtered$depth)". I changed the code a bit and the result is <-0.8. I finally asked Bard what it meant and it explained to me."An R-squared value of 0.8 or higher suggests a good fit, while an R-squared value of 0.6 or lower suggests a weaker fit." here is the link to the chat https://g.co/bard/share/a36eb7e61330

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

```
#7
##performing a lm model to see what the relationship is between depth and temp
#cor(NTL_filtered$temperature_C, NTL_filtered$depth)

NTL_TD <- NTL_filtered %>%
  select(depth, temperature_C)

# Fit a linear regression model to the data
lm_model <- lm(temperature_C ~ depth, data = NTL_TD)

# Display the results of the linear regression model
lm_model
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = NTL_TD)
##
## Coefficients:
## (Intercept)        depth
##      21.956       -1.946
```

```
summary(lm_model)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = NTL_TD)
##
## 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
```

```
####works cited: Bard -- https://g.co/bard/share/a36eb7e61330.
#Prompt: Perform a linear regression to test the relationship
#and display the results
```

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: how much of the variability in temperature is explained by changes in depth: this is the R-squared value –> 0.74 –> 74% The degrees of freedom on which this finding is based: 9726 degrees of freedom (from the summary) The statistical significance of the result: This depends on the size of the p-value. The P value is <2e-16, which is very small –> means that the statistical significance of the result is very high How much temperature is predicted to change for every 1m change in depth: can be seen from the coefficient -1.94, which means that for for 1m change in depth, the temp decreases by 1.94 degrees C. Works cited: Bard (for "Also mention how much temperature is predicted to change for every 1m change in depth") I fed it the result and used the prompt "how much temperature is predicted to change for every 1m change 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
#summary(NTL_filtered$lakename)
### running an AIC to see what explantory vairable sets best explains temp
TPAIC <- lm(data = NTL_filtered, temperature_C ~ depth + daynum, year4)
step(TPAIC)
```

```
## Start:  AIC=15882.02
## temperature_C ~ depth + daynum
```

```
##
##          Df Sum of Sq    RSS   AIC
## <none>                  49750 15882
## - daynum  1     14014  63764 18294
## - depth   1    252608 302357 33435

##
## Call:
## lm(formula = temperature_C ~ depth + daynum, data = NTL_filtered,
##     subset = year4)
##
## Coefficients:
## (Intercept)        depth       daynum
##    -412.680       -1.717        2.372
```

```
#10
## running a multiple regression on the recommended set of variables
TPmodel <- lm(formula = temperature_C ~ depth + daynum, data = NTL_filtered,
    subset = year4)
summary(TPmodel)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + daynum, data = NTL_filtered,
##     subset = year4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7396 -1.6555 -0.0423  1.6835  4.8091
##
## Coefficients:
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept) -412.679702    8.295019  -49.75   <2e-16 ***
## depth         -1.717123    0.007727 -222.22   <2e-16 ***
## daynum         2.371705    0.045314   52.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.262 on 9725 degrees of freedom
## Multiple R-squared:  0.8875, Adjusted R-squared:  0.8874
## F-statistic: 3.835e+04 on 2 and 9725 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 set of explanatory variables: depth + daynum, and also subsetting the values to their repective years. "Call: lm(formula = temperature_C ~ depth + daynum, data = NTL_filtered, subset = year4)" This model explains 88.7% of the observed variance. "Multiple R-squared: 0.8875, Adjusted R-squared: 0.8874" This is indeed an improvement from only using depth as the predictor, which had an R squared value of 74%.

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
###Model that tests whether the different lakes have, on average,
#different temperatures in the month of July expressed as an ANOVA test
Lake.differences.anova1 <- aov(data = NTL_filtered, temperature_C ~ lakename)
summary(Lake.differences.anova1)
```

```
##               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
```

```
###Model that tests whether the different lakes have, on average,
#different temperatures in the month of July expressed as a linear model
Lake.differences.anova2 <- lm(data = NTL_filtered, temperature_C ~ lakename)
summary(Lake.differences.anova2)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = NTL_filtered)
##
## 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: There is a significant difference in mean temperature among the lakes. The P value is extremely small(<2e-16), which means that the null hypothesis of the lake mean temperature being the same is rejected. As a result, we know that the means across different lakes are not the same/ the difference between group means is statistically significant.

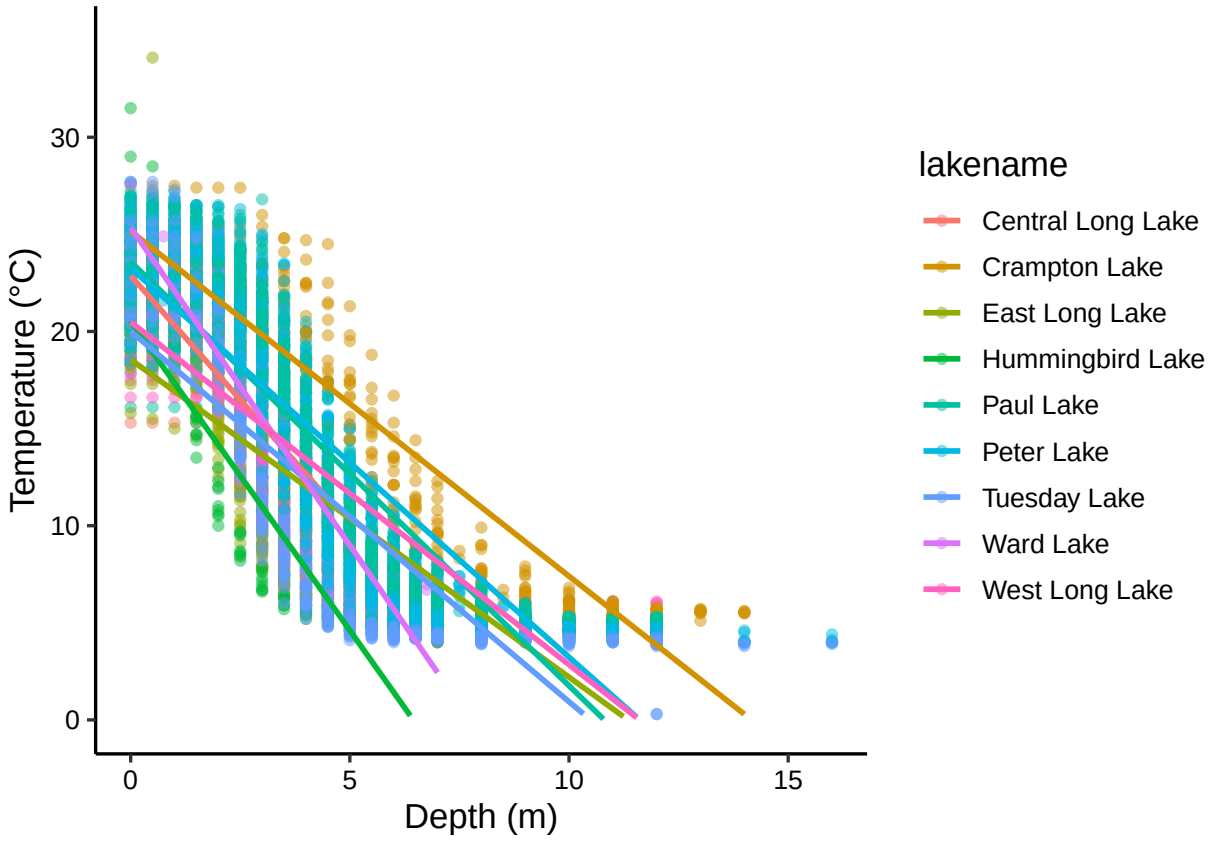
14. Create a graph that depicts temperature by depth, with a separate color for each lake. Add a geom_smooth (method = "lm", se = FALSE) for each lake. Make your points 50 % transparent. Adjust your y axis limits to go from 0 to 35 degrees. Clean up your graph to make it pretty.

```
#14.
Lake.differences.plot <- ggplot(NTL_filtered, aes(x = depth,y = temperature_C,
                                                  color =lakename)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  ylim(0,35) +
  mytheme +
  labs(x = "Depth (m)", y = "Temperature (°C)")

print(Lake.differences.plot)
```

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

```
## Warning: Removed 73 rows containing missing values (`geom_smooth()`).
```



```
#works cited: Bard. Wrote code on my own and updated it with some details in the code from Bard
```

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

```
#15
#running Tukey Honest Significant Differences.
TukeyHSD(Lake.differences.anova1)
```

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

```
Lake.differences.groups <- HSD.test(Lake.differences.anova1, "lakename",
                                    group = TRUE)
Lake.differences.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"
```

```
### accordign to the HSD.test,
### Central Long Lake and Crampton Lake,
### Campton lake and Ward Lake,
### Ward Lake, Peter lake, and west long lake,
### west long lake, Tuesday laek, and hummingbird lake,
### and Tuesday laek,  hummingbird lake, and east long lake have similar means.
### in other words, the other pairs would have different means.
```

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 and Paul Lake would have the same mean temperature as Peter Lake. All lakes have at least one lake that has a similar/same mean temperature as itself (as represented by the groups).

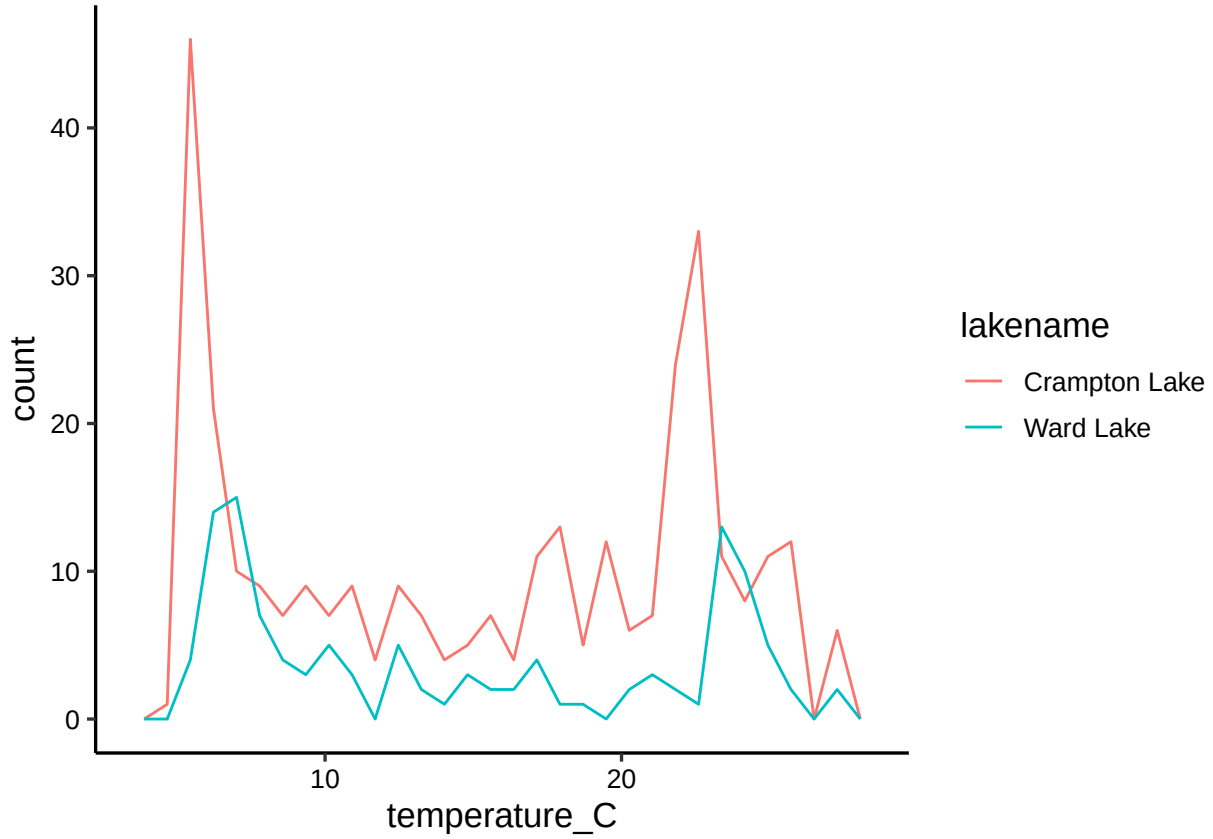
17. If we were just looking at Peter Lake and Paul Lake. What's another test we might explore to see whether they have distinct mean temperatures?

> Answer: Given the Q25 and Q50 of the two lakes vary by around 1, it would be interesting to explore the distribution of the two lakes. To do that, we can employ the Kolmogorov-Smirnov test to compare the cumulative distribtuion functions of the two lakes. ####works cited: Bard.Prompt: "what is a test, other than the tukey test, that can be used to explore two lakes with the same mean temperature (as given by a tukey test), whether they have distinct mean temperatures?" It gave me several options, and i chose the one that is most suitable for Paul and Peter lakes. A two sample t-test can also be used to see if the two means are the same.

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?

```
##filtering the data so that it only includes data for Crampton and Ward lakes
NTL_CW <- NTL_filtered %>%
  filter(lakename %in% c("Crampton Lake", "Ward Lake"))
##visualizing the data
ggplot(NTL_CW, aes(x = temperature_C, color = lakename)) +
  geom_freqpoly()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
### the two have very different distributions
CW.twosample <- t.test(NTL_CW$temperature_C ~ NTL_CW$lakename)
CW.twosample
```

```
##
##  Welch Two Sample t-test
##
## data:  NTL_CW$temperature_C by NTL_CW$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
```

```
#### p-value = 0.2649, mean for CL: 15.35, mean for WL: 14.46
### the means are the same.
```

Answer: The p-value of 0.26 means taht we fail to reject the null hypothesis, and that there is no statistically significant difference between the two groups–> the means for the groups are the same. The mean temperatures are similar, being 15.35 for Crampton and 14.46 for Ward. This result matches my answer for part 16 that these two have the same means.